ORIGIN AND EVOLUTION OF PROTEIN FOLD DESIGNS

BY

SYED ABBAS BUKHARI

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Bioinformatics
with a concentration in Crop Science
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2012

Urbana, Illinois

Adviser:

Professor Gustavo Caetano-Anolles

ABSTRACT

The spatial arrangements of secondary structures in proteins, irrespective of their connectivity, depict the overall shape and organization of protein domains. These features have been used in the CATH and SCOP classifications to hierarchically partition fold space and define the architectural make up of proteins. Here we use phylogenomic methods and a census of CATH structures in hundreds of genomes to study the origin and diversification of protein architectures (A) and their associated topologies (T) and superfamilies (H). Phylogenies that describe the evolution of domain structures and proteomes were reconstructed from the structural census and used to generate timelines of domain discovery. Phylogenies of CATH domains at T and H levels of structural abstraction and associated chronologies revealed patterns of reductive evolution, the early rise of Archaea, three epochs in the evolution of the protein world, and patterns of structural sharing between Archaea and Eukarya that are very recent. Trees of proteomes confirmed the early appearance of Archaea in the world of organisms. Phylogenies reconstructed from phylogenetic character sets representing T and H domains of different age congruently reflected patterns of domain appearance in the structural chronologies. Trees reconstructed from ancient domain revealed an archaeal rooting. In contrast, trees reconstructed from modern domains exhibited the canonical bacterial rooting. Timelines suggest this rooting is probably driven by patterns of sharing between Archaea and Eukarya. Although CATH and SCOP differ significantly in domain definitions, our findings indicate both classification schemes apportion protein structures on very similar theoretical grounds that harbor similar phylogenetic history. Phylogenies of CATH domains at A level of structural abstraction uncovered general patterns of architectural origin and diversification. The tree of A structures showed that the *3-layer (αβα) sandwich (3.40)* and the *orthogonal bundle (1.10)* that harbor simple secondary

structure arrangements are the most ancient, popular and abundant structural designs of proteins. Phylogenies also revealed that ancient A structural designs are comparatively simpler in their makeup and are involved in basic cellular functions. In contrast, modern structural designs such as *prisms, propellers, 2-solenoid, super-roll, clam, trefoil and box* are not widely distributed and were probably adopted to perform specialized functions. Our timelines therefore uncover a universal tendency towards protein structural complexity that is remarkable.

ACKNOWLEDGMENTS

I would especially like to thank Dr. Gustavo Caetano-Anolles for acting as my adviser, and for his committed guidance and assistance during the research and preparation of my thesis. I would also like to thank Dr. James Whitfield and Dr. Sandra Rodriguez-Zas for serving on my graduate committee and providing helpful suggestions and comments on this project. I also wish to thank my fiancé Bushra Fazal Minhas for helping me compiling the thesis and suggestions. Thanks are also extended to my lab fellows, Dr. Jay, Dr. Minglei Wang, Fayyez Aziz and Arshan Nasir for suggestions and recommendations.

TABLE OF CONTENTS

CHAPTER 1

Literature Review: Evolutionary Origins Of "Protein Fold Space" And Prevailing Approaches Towards Protein Structure Classification

"*Possibly the most pregnant recent development in molecular biology is the realization that the beginnings of life are closely associated with the interactions of proteins and nucleic acids.*"

— Florence O. Bell, "X-ray and Related Studies of the Structure of the Proteins and Nucleic Acids", Leeds PhD Thesis (1939), quoted in Robert Olby, The Path to the Double Helix: The Discovery of DNA (1994).

1.1 INTRODUCTION

The totality of proteins in all organisms on Earth is vastly and collectively referred as the "*protein universe*". Rossmann and Argos (1976) systematically compared the protein structures and noted that in fact molecular structure seemed to be more conserved than sequence, particularly around their binding sites. This helped establish the most accepted notion among structural biologists that "*That protein structure is more conserved than sequence*". Analyses of known structures have suggested that the protein structural universe is redundant and the total number of possible folds is limited. Understanding the nature of protein structural space can help characterize the relationship between protein sequence, structure and functions.

Considerable effort has been made to study the structural space of proteins, the most notable one, the Protein Data-Bank (PDB), was established to collect the growing set of protein structural models (Bernstein et al. 1977). Several classification schemes were proposed to organize this set of structures, but were focused on specific group of folds. Classifiers that were general to all proteins were introduced later, such as secondary structure elements and domains (Richardson 1981). However these initial developments for classifying protein structures provided a fundamental framework for development of more comprehensive and maintained databases a decade later (Holm et al. 1992; Murzin et al. 1995; Orengo et al. 1997; Holm and Sander 1998). These databases were aimed to reflect evolutionary as well as functional relationships among protein folds. Many of the classification rules are based on a mixture of concepts. Ideally, these classification systems should reflect evolutionary history up to the highest level.

Modern physical techniques, such as high-throughput synchrotron-based X-ray crystallography and multi-dimensional NMR, promise rapid growth in the number of known

protein structures (Heinemann et al. 2001; Liu et al. 2007; Service, 2008). As of today (May 25, 2012), the PDB contains 81,756 known protein structures and the average number of structures acquired per year for last five years (2007-2011) is more than 7000. Despite this flood of data, increasingly efficient and robust methods for protein three-dimensional (3D) structural comparisons make it feasible to perform all-against-all comparisons of all known structures. These structural comparison methods revealed that proteins can share a common fold despite little or no sequence similarity. Similarly, proteins with a same fold can have different functions. However the main aim is to try to bring some order into the description of protein structure by imposing a classification (Taylor and Aszodi 2005). The most popular classifications accessible via the World-Wide Web (WWW) are: (a) SCOP: a "structural classification of proteins" which is essentially a manual classification, (b) CATH: a semi-automated system which uses both manual and automated approaches, and (c) FSSP: a 3D alignment system that uses the DALI program and is totally automated. SCOP and CATH are the most accepted classification schemes today.

SUPERFAMILY (http://supfam.cs.bris.ac.uk/SUPERFAMILY/) and Gene3D (http://gene3d.biochem.ucl.ac.uk/Gene3D/) are the databases of fold recognition assignments to fully sequenced genomes developed based on SCOP and CATH classification systems respectively.

SCOP

The first comprehensive classification of PDB structures to be made available on the web was SCOP. The SCOP (Structural Classification Of Proteins) database was established to infer relationship between protein structure and sequence based on current available data (Murzin et al. 1995). This database contains complete classification of all the available proteins in PDB

along with some additional structures that have not been deposited yet. SCOP makes use of the structure, function and evolutionary origin of protein domains to organize them into a hierarchy. The unit of categorization in the hierarchy is the domain, since domains are typically the units of protein evolution, protein structure, and molecular function. SCOP is based on a four-level hierarchy, the top one being protein 'class' in which folds are grouped into four major general definitions: all-α, all-β, α/β,and α+β along with others that deal with membrane associated proteins and peptides. The second level is the 'fold' level, which describes the topology of proteins a common core structure. In the third level, folds are divided into 'superfamilies', which consist of proteins that are thought to be evolutionary related. The final general level in the SCOP hierarchy is the 'family' level that contains protein domains having high sequence identity. This indicates a close evolutionary relationship.

CATH

The CATH Protein Structure Classification is a semi-automatic, hierarchical classification of protein domains proposed by Christine Orengo, Janet Thornton and their colleagues (Orengo et al. 1997). The authors of CATH attempt to automate their classification process as much as possible without losing biological relevance. In addition, the structural aspects of classification are weighted more in CATH. To reflect this, CATH has an additional hierarchical level (Architecture level) when compared to SCOP (Figure 1.1).

The name CATH is an acronym of the four main levels of its classification: (1) **C**lass (C): the overall secondary-structure content of the domain; (2) **A**rchitecture (A): high structural similarity but no evidence of homology; (3) **T**opology (T): a large-scale grouping of topologies that share particular structural features; and (4) **H**omologous superfamily (H): indicative of a

demonstrable evolutionary relationship. Here topology is similar to SCOP's fold level, and homology is similar to SCOP's superfamily level. The 'Architecture' level is specific to CATH and represents the shape defined by the assembly of secondary structures without considering their connectivity.

CATH uses several computational tools to facilitate classification of new PDB entries. This can be summarized into four major steps: (1) filtering of low resolution structures; (2) sequences are matched against the domains already have been classified; (3) structural comparison is made to identify new potential structures or found previously; and, (4) Finally, new PDB entries are split into sequence families, underneath the homology level. CATH uses SIFT protocol filtering (Michie et al. 1996) and consider crystal structures solved to resolution better than 4.0 Å along with NMR structures. All non-proteins, models, and structures with greater than 30% "C-alpha only" are excluded from CATH. The classification is performed on individual protein domains.

To divide multidomain protein structures into their constituent domains, a combination of automatic and manual techniques are used. If a given protein chain has sufficiently high sequence identity and structural similarity (i.e. 80% sequence identity, SSAP score $\geq$ 80) with a chain that has previously been chopped, the domain boundary assignment is performed automatically by inheriting the boundaries from the other chain. Otherwise, the domain boundaries are assigned manually, based on an analysis of results derived from a range of algorithms which include structure based methods (CATHEDRAL, SSAP, DETECTIVE (Swindells 1995), PUU (Holm & Sander 1994), DOMAK (Siddiqui and Barton 1995), sequence based methods (Profile HMMs) and relevant literature (Greene et al. 2007).

FSSP and Dali-domain server

The Family of Structurally Similar Proteins (FSSP) is a database of structurally superimposed proteins generated using the "Distance-matrix ALIgnment" (DALI) algorithm (Holm et al. 1992). FSSP classification is a structure alone system; protein sequence and function are not taken into account. The database is helpful for the comparison of protein structures. Although the FSSP is not a hierarchical classification, it clusters similar structures together into a tree of folds so that it is easy to analyze a particular family.

SUPERFAMILY

SUPERFAMILY classifies amino acid sequences into known structural domains on completed genomes, especially into SCOP superfamilies (Gough and Chothia 2002). Protein sequences from completely sequenced genomes are scanned against hidden Markov models (HMMs), using an automated fold recognition procedure SAM-T90 that was fine tuned with expert knowledge to recognize superfamilies as defined by SCOP. For each sequence in a superfamily, filtered at 95% sequence identity, HMMs were built using homologues from a non-redundant sequence database. The HMMs for each seed were then scored against the set of genes from the completed genomes. Interestingly, this procedure gave better results than forming a single HMM from a structural alignment of all sequences in a SCOP superfamily (Madera and Gough 2002). In the SAM-T99 procedure the reversed score of the search sequence normalizes e-values.

Gene3D

Gene3D is a resource similar to SUPERFAMILY that assigns CATH domains rather than SCOP domains to completed genomes. To save computational time, the genes on the genomes are first clustered into families based solely on sequence information. HMMs are built for CATH domains using the SAM-T99 technology. The HMMs are then scored against representative sequences from the family clusters, filtered at 35% sequence identity (Lee et al. 2005). The final domains are assigned through a 'Domain Finder' method, which checks for significance (e-values) and overlap.

A "Periodic Table" perspective of protein structural space

As discussed above, the systematic classification of protein structural space is an essential exercise promising to organize knowledge and support new hypotheses about the physics and evolution of protein structure. Both SCOP and CATH are now considered standard tools for benchmarking structure prediction and evolutionary inference. However, both structural classification schemes define protein fold structure differently. Several studies have shown a number of difficulties with these hierarchical approaches to the organization of structural data, invoking the need for alternative strategies Cuff et al, 2009; Harrison et al, 2002; Reeves et al, 2006). An alternative view of protein structure space can be found by using topological descriptions that cover large and well-folded structures and are defined by a "periodic table" of protein structure. The periodic table defines entities as ideal forms, which identify regions conforming to the overall form architecture of proteins. Taylor (2002) defined ideal forms that cover arrangements of β–strands and α–helices in three-layer, four-layer and barrel organizations. However the elements defined in the periodic table and the rules governing the

transition of one topology into another were not elaborated within an evolutionary framework, questioning the appropriateness of the method.

1.2 THESIS RESEARCH MOTIVATION

The shapes of proteins (generally referred as protein fold) are not only the result of their history but also of the physiochemical constraints (e.g. the strength of covalent and hydrogen bond interactions), the environment in which they operate (e.g., aqueous, lipid, intracellular, extracellular) and their functional role (e.g., catalysis, signaling). It is a very difficult task to separate these constraints from those that are inherited (Taylor and Aszodi 2005).

It is therefore fundamental that we dissect evolutionary and physiological components to better understand the evolution of redundant protein topologies. One approach is to study the evolutionary appearance (i.e. assign an age) and the distribution and diversification of structural designs (i.e. study how widely distributed are architectures) by focusing on the conservation of protein structures across lineages. The repertoire of protein structures encoded in genomes is evolutionarily conserved and capable of preserving an accurate record of genomic history (Caetano-Anollés and Caetano-Anollés 2003; Wang et al. 2007, 2011). A considerable number of studies have been conducted to unfold the evolutionary mechanism of protein domain distribution and evolution in the world of organisms we see today (Wang et al. 2007, 2011). Here we explore the appearance and diversification of general protein structural designs, such as *sandwiches*, *bundles*, *barrels*, *solenoids*, *ribbons*, *prisms*, *propellers* and *trefoils*. These designs are defined by the CATH (Orengo et al. 1997) and SCOP (Murzin et al. 1995) protein structural classifications. We focus on building trees of structures instead of creating universal organismal trees at sequence level, making our approach quite unique and innovative.

Figures

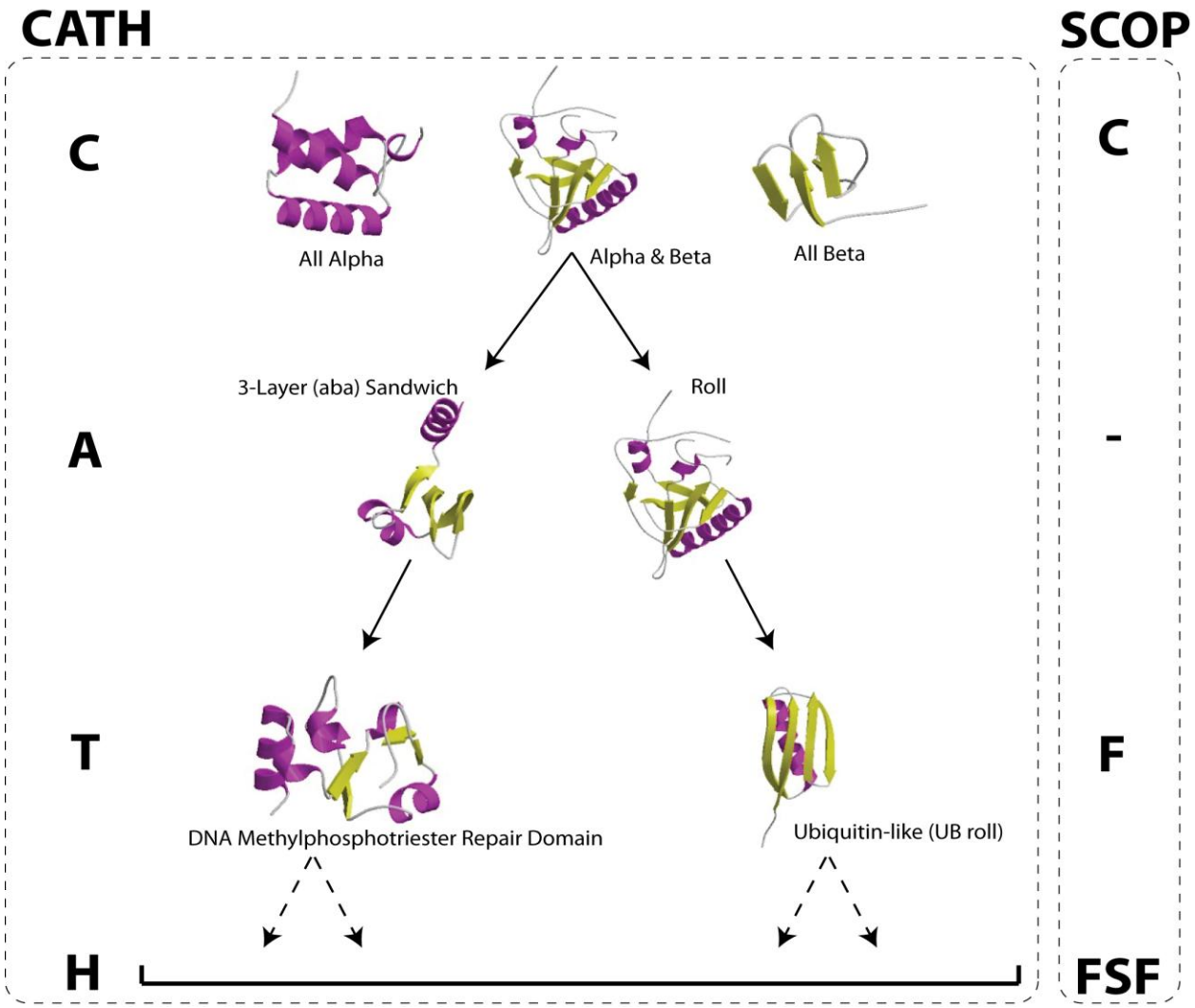

Figure 1.1 This figure describes the hierarchy of the CATH structural classification system and also shows corresponding SCOP levels. Architecture (A) level is unique to CATH structure classification system.

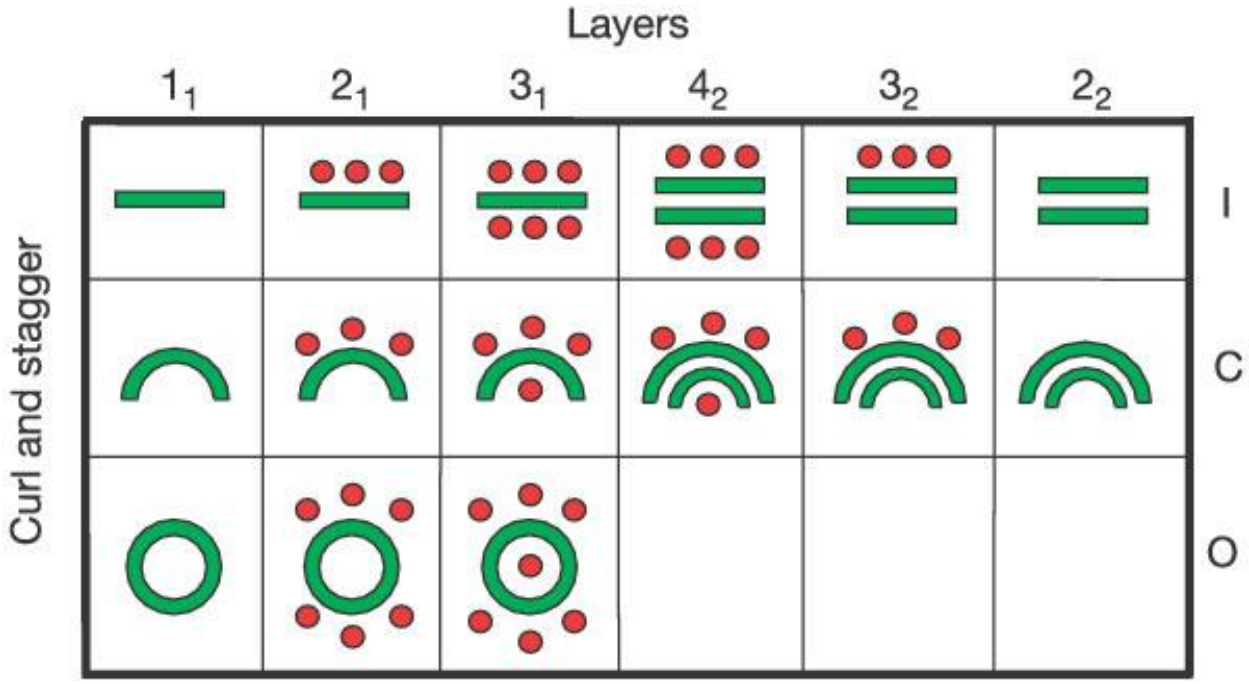

Figure 1.2 A periodic table of protein structures describes the simplified layer structure of proteins. Alpha (red) and beta (green) layers of protein secondary structures are combined to make globular domains. I, C and O represent flat, curled and barrel respectively. B-sheets normally have a twist, which can result in whole structure twist, allowing them to adopt curl, which can incorporate a stagger between adjacent strands. Combinations of these curl and stagger results in barrel structure. The top axis depicts the layer combination of both alpha and beta (subscript) secondary structures [Taken From Taylor, 2002].

CHAPTER 2

Origin Of Protein Fold Designs, Modern Archaeo-Eukaryotic Architectural Sharing And Proteome Evolution Inferred From Phylogenomic Analysis Of CATH Domain Structures

"Perhaps the most remarkable features of the [myoglobin] molecule are its complexity and lack of symmetry. The arrangement seems to be almost totally lacking in the kind of regularities which one instinctively anticipates, and its more complicated than any theory of protein structure."

— John Kendrew et al (1958)

2.1 INTRODUCTION

The polypeptide chains of proteins generally fold into highly ordered and well-packed 3D atomic structures (Caetano-Anolles et al. 2009). These protein folds represent spatial arrangements of more or less wound helices (generally α–helices) and extended chain segments (β–strands) that are separated by relatively rigid loop regions in the form of turns and coils. Helices are stabilized by local main-chain (backbone) hydrogen bonding interactions. In turn, β–strands establish main-chain interactions with other strand elements that are distant. Parallel and antiparallel arrangements of β–strands form β–sheets, which often curve to form open and closed barrel structures. Folds are generally defined by the composition, architecture and topology of their core 'helix' and 'sheet' secondary structure elements (Andreeva and Murzin 2006). The satisfaction of the hydrogen bonding potential of main-chains gives rise to regular secondary and super secondary structural elements in globular proteins. Analysis of protein folds indicates that those that occur frequently tend to adopt regular architectures, such as the αβ–Rossmann folds, αβ–barrels, β–sandwiches, and bundles (Worth et al. 2009). Main-chain hydrogen bonding is also important for the formation of complex turns and coils that link α–helices and β–strands.

Protein domains are compact, recurrent, and independent folding units of protein structure that sometime combine with other domains to form multidomain proteins. They are considered evolutionary units and are the basis for several protein structure classification schemes. Two of them, CATH and SCOP, are accepted as gold standards and share a number of common features (Csaba et al. 2009). SCOP (Murzin et al. 1995) is a largely manual collection of protein structural domains that aims to provide a detailed and comprehensive description of

the structural and evolutionary relationships of proteins with known structures. In contrast, CATH (Orengo et al. 1997) uses a combination of automated and manual techniques, which include computational algorithms, empirical and statistical evidence, literature review and expert analysis. Both classifications are hierarchical but dissect 3D structure differently, focusing more on either evolutionary or structural considerations (Csaba et al. 2009). SCOP unifies domain structures that are evolutionarily related at sequence level (>30% pairwise residue identities) and are unambiguously linked to specific molecular functions into fold families (FFs), FFs with common structures and functions share a common evolutionary origin into fold superfamilies (FSFs), FSFs with similarly arranged and topologically connected secondary structures (not always evolutionarily related) into folds (Fs), and finally Fs that share a general type of structure into classes. CATH unifies domain structures hierarchically (bottom-up) into sequence families (SFs; analogous to FFs), homology superfamilies (Hs; analogous to FSFs), topologies (Ts; analogous to Fs), architectures (As), and protein classes (Orengo et al. 1997) (see also Figure 1.1 for comparisons of SCOP and CATH levels of structural abstractions.). Multi-linkage clustering groups domains into SFs based on sequence similarity. SFs with structures that are thought to share common ancestry and can be described as homologous are grouped into Hs. H structures sharing patterns of overall shape and connectivity of secondary structures are grouped into Ts. T structures that share and overall shape of the domain structure according to the orientations of the secondary structures but ignoring their connectivity are unified into As. Finally, A general shapes are grouped into four protein structural classes, mainly-alpha, mainly-beta, alpha-beta and few secondary structures (Orengo et al. 1997).

Protein structures are evolutionarily conserved and capable of preserving an accurate record of genomic history (Wang et al. 2007; Caetano-Anolles et al. 2009). They represent

'living fossils' of molecular evolution (Andreeva and Murzin 2006) and express the greatest levels of redundancy and reuse that exist in molecular biology (Gerstein et al. 1998). Many studies have been conducted to unfold the evolution and diversification of protein domain structures and proteomes of extant organisms (Caetano-Anolles et al. 2009; Chothia et al. 2009; Forsland et al. 2008; Kim et al. 2011). Structural phylogenies describing the evolutionary relationship of SCOP F, FSF and FF domains were built by data-mining a census of structures in hundreds of genomes (Caetano-Anollés and Caetano-Anollés 2003; Wang et al. 2009; Bussemaker et al. 1997; Caetano-Anollés et al. 2011). Timelines of F, FSF and FF appearance were derived from the phylogenetic trees and revealed the existence of three epochs in protein evolution, architectural diversification, superkingdom specification and organismal diversification. Patterns of reductive evolution in the domain repertoire unfolded in the timelines consistently segregated the archaeal lineage from the ancient community of organisms and established a first organismal divide during the superkingdom specification epoch. Finally, trees of proteomes (i.e. trees of life) placed Archaea at the root and confirmed this organismal supergroup represents the most ancient superkingdom of life (Wang et al. 2007; Kim and Caetano-Anollés 2012).

While we have studied how F, FSF and FF domains appeared and distributed in the world of organisms, we have not embarked in a systematic study of the origin and evolution of general structural motifs. Here we study how structural designs evolve in trees of domain structures, this time focusing on the CATH classification. The appearance and diversification of general protein structural designs at A-level (e.g., *sandwiches, bundles, barrels, solenoids, propellers etc.*) and published literature define in this study a unique chronology of structural innovation. Structural phylogenies of domains at T and H levels of structural abstraction uncover global evolutionary

patterns of structural distribution in the world of organisms. The study benchmarks previous phylogenetic analysis of SCOP-defined domains and again reveals the early origin of the archaeal superkingdom. Congruent patterns of diversification derived from protein structure provide remarkable support to the ancient history of the cellular world, and trees of life confirm the primordial evolutionary patterns.

2.2 MATERIALS AND METHODS

Phylogenomic trees describing the evolution of domain structures and proteomes were reconstructed using a census of domain abundance in proteomes using PAUP* version 4.0b10 (Swofford 2002). Figure 2.1 presents the flowchart of the adopted methodology. CATH annotations for the proteomes of 492 fully sequenced genomes (42 Archaea, 360 Bacteria and 90 Eukarya) were retrieved from *Gene3D* (Lees et al. 2009). Table S1 lists the organisms studied. *Gene3D* is a repository of manually curated HMM predictions with a false positive prediction rate of only 0.2-0.6%. As with *SUPERFAMILY* (Gough and Chothia, 2002; and Chothia et al. 2009), a repository of SCOP domain predictions, proteomes deposited in *Gene3D* were searched against HMM libraries using the iterative Sequence Alignment and Modeling System (SAM) method. Data matrices of genomic abundance (*G*) of domains at A, T and H levels were assembled for phylogenetic analysis. Empirically, *G* values represent numbers of multiple occurrences of an A, T and H domain in a genome, ranging from 0 to thousands and resembling morphometric data with large variances. Because existing phylogenetic programs can process only tens of phylogenetic character states depending on user's CPU performance, the space of *G* values in the matrix was reduced using a standard gap-coding technique with the following formula:

$$G_{xy\_norm} = \text{Round}\left[\frac{\ln(G_{xy}+1)}{\ln(G_{xy\_max}+1)} \times 20\right]$$

in which x and y denote an A, T or H domain structure, y a genome, and G_{xy} the abundance of x in y. G_{xy_max} indicate maximum G_{xy} values for all y genomes. The round function normalizes *G* values on a 0-20 scale (G_{xy_norm}). These values define character states, which are encoded as linearly ordered multistate phylogenetic characters using an alphanumeric format of numbers 0–9 and letters A–K that is compatible with PAUP*. Transposition of the data matrix (switching characters and taxa) allowed reconstruction of trees of either proteomes or domain structures. Trees of A, T and H domains were built by polarizing states from 'K' to '0' using the ANCSTATES command in PAUP*, with 'K' being ancestral. Trees of proteomes were built by polarizing character states from '0' to 'K', with '0' being ancestral. The trees were rooted without invoking outgroup taxa using the Lundberg method, which positions the most ancient proteomes and domain structures at the base of their corresponding trees. Assumptions of character argumentation have been discussed in previous publications (Caetano-Anollés and Caetano-Anollés 2003; Wang et al. 2007; Caetano-Anolles et al. 2009; Caetano-Anollés et al. 2011). Our model of structural evolution ('K' to '0' polarization) considers that the abundance of individual domain structures increases progressively in nature, even when expanding domain levels suffer loss in individual lineages or are selectively constrained during evolution (we consider that character state transformation is reversible). Consequently, ancient structures are more abundant and widely present than younger ones. In contrast, our model of proteome evolution ('0' to 'K' polarization) assumes proteomes have built their structural repertoires progressively, increasing both the diversity and abundance of their structural make up. Consequently, genomes that are ancient developed their repertoires earlier from a pool of

structures that was comparatively simpler. Their repertoires are today simpler than those that developed their repertoires more recently from a more complex and diverse pool of structures.

Phylogenomic trees were reconstructed using the maximum parsimony (MP) optimality criterion in PAUP* with 1,000 replicates of random taxon addition, tree bisection reconnection (TBR) branch swapping, and maxtrees unrestricted. Phylogenetic confidence was evaluated by the nonparametric bootstrap method with 1,000 replicates (resampling size matches the number of the genomes sampled; TBR; maxtrees, unrestricted). The degree of phylogenetic signal for taxa was measured using the skewness (g_1) test with a tree length distribution obtained from 1,000 random trees.

Since trees of domain structures are rooted and are highly unbalanced, we unfolded the relative age of protein domains directly for each phylogeny as a distance in nodes (node distance, *nd*) from the hypothetical ancestral architecture at the base of the trees in a relative 0–1 scale. *nd* was calculated by counting the number of internal nodes along a lineage from the root to a terminal node (a leaf) of the tree on a relative 0–1 scale with the following formula:

$$nd_a = \frac{No.\,of\ internal\ nodes\ between\ nodes\ r\ and\ a}{No.\,of\ internal\ nodes\ between\ nodes\ r\ and\ m}$$

where *a* represents a target leaf node (either an A, T or H domain), *r* is a hypothetical root node, and *m* is a leaf node that has the largest possible number of internal nodes from node *r*. Consequently, the *nd* value of the most ancestral taxon is 0, whereas that of the most recent one is 1. Node distance can be a good measure of age given a rooted tree because the emergence of protein domains (i.e., taxa) is displayed by their ability to diverge (cladogenesis or molecular

speciation) rather than by the amount of character state change that exists in branches of the tree (branch lengths).

2.3 RESULTS AND DISCUSSIONS

Structural chronologies of CATH domain structures uncover patterns of proteome diversification

We generated phylogenomic trees describing the phylogenetic relationship of 38 A, 1,152 T and 2,221 H domain structures (Figure 2.2 and 2.9). Tree distribution profiles and metrics of skewness suggested significant cladistics support ($P < 0.01$). The trees were well resolved. However, internal branches were poorly supported by bootstrap analysis, an expected outcome with trees of this size. Chronologies of evolutionary appearance (Wang et al. 2007) of CATH domain structures were derived directly from the phylogenomic reconstructions (Figure 2.2 and 2.9). The relative age of domains (*nd*) was measured on the trees as a relative distance in nodes from the hypothetical ancestor of domains at the base of the trees, and used to build the timelines. Since our method produces rooted trees that are highly unbalanced and reject the Yule and random speciation models (Steel and McKenzie 2001) and since molecular speciation in our trees has clock-like behavior and does not depend on changes in domain abundance (Wang et al. 2011), *nd* was considered a good and most-parsimonious proxy for time. To study how domain structures distribute in proteomes, we calculated a *distribution index* (*f*), the number of species that use each structure given on a relative 0-1 scale. The *f* index was plotted along the timelines of domain structures, i.e. against *nd*. Three As (nd_A = 0-0.068), fifteen Ts (nd_T = 0-0.061) and fifteen Hs (nd_H = 0-0.049) were present in all proteomes we examined (f = 1) and were the most ancient in the timeline. The *f* of Ts and Hs decreased with their increasing age until *f* approached

zero at nd_T = 0.55 and at nd_H = 0.55, respectively. We term these ages "crystallization points" of the T and H structural chronologies, borrowing the idea of a phase transition from physics. At these time points, a steady decrease in *f* results in a large number of structures being specific to a small number of organisms. After crystallization, an opposite trend takes place, in which Ts and Hs increase their representation in genomes. In contrast, the architectural chronology that describes the appearance of As remained unaffected by the crystallization event since the loosing trend of As started at nd_A = 0.56-0.60 but rarely reached zero (see below).

To uncover hidden patterns of organism diversification in our dataset, we divided structures according to their distribution in superkingdoms Archaea (A), Bacteria (B) and Eukarya (E) and constructed three separate structural chronologies for the genomes of each superkingdom at A, T and H levels of structural abstraction (Figs. 2-3, 2-4 and 2-10). Domain structures were pooled into seven taxonomical groups depending on whether they were unique to a superkingdom (A, B and E) or shared by two (AB, AE and BE) or three superkingdoms (ABE). Taxonomical groups were identified in the time plots with different colors. We previously observed that a superkingdom must 'lose' a significant number of SCOP structures before the first superkingdom-specific 'signature' structure appeared in evolution (Wang et al. 2007). In our study, this loser trend of domain structures was also observed for the CATH annotated genomes in each superkingdom. This observation strengthens our claim of reductive evolution in the domain content of the lineages that emerge from the common ancestor [the 'urancestor' or the Last Universal Common Ancestor (LUCA)] that we find is functionally complex (Kim and Caetano-Anollés 2011). The loser trend of SCOP and CATH structures reveals the primordial birth of Archaea followed by the birth of Bacteria and Eukarya. The complete loss of Hs first starts in Archaea (nd_H = 0.176) with the membrane-bound lytic murein transglycosylase D (chain

A) H domain (*3.10.350.10*). Its appearance is congruent with the loss of the first SCOP FSF in Archaea (nd_{FSF} = 0.174), the *LysM domain* (d.7.1), observed in previous studies (Wang et al. 2007). Both domain definitions are very much similar in how they describe functions in the cell. Analysis of domain distribution in Archaea shows that the vast majority of ancient Ts and Hs that were lost in proteomes were present in all superkingdoms (ABE; colored grey). These were followed by AB (orange), A (wine) and few AE (red) structures, most of which started to appear after the crystallization point and during the superkingdom specification and organismal diversification epochs. Clear decreases in structural representation (*f*-value) also occurred in Bacteria and Eukarya, but involved fewer and younger structures. Analysis of domain distribution in Bacteria shows that AB and B structures (dark yellow) started to increase representation after the crystallization point, leading towards their diversification and specification. Similarly, the eukaryotic chronology showed that comparatively younger architectures [e.g. BE (blue) and E (green)] increased their popularity among the eukaryal lineages. The appearance and distribution of the seven taxonomical groups of H and T structures was unfolded in the timelines using boxplots describing the range of nd_H and nd_T values and measures of central tendency for each group (Figure 2.5). Only domains shared by the three superkingdoms (ABE) span the entire chronology, from the origin of proteins (nd = 0) to the present (nd = 1). These structures represent instantiations of the domain content of LUCA but their late appearance may also indicate events of horizontal transfer between lineages. Boxplots for BE, AE and AB explain relationships among superkingdoms over time. The BE boxplot is the most ancient of the three, suggesting Archaea diversified early by reductive evolution. The A, B and E boxplots reflect the history of 'signature' structures that are unique to individual superkingdoms. These signatures appear first in Bacteria and then concurrently in Archaea and

Eukarya, an observation that is congruent with timelines derived from SCOP domains (Wang et al. 2007). Despite its early specification, Archaea tends to acquire Archaea-specific structures very late in evolution and their number is limited when compared to Bacteria and Eukarya. This may stem from very strong adaptive pressures historically imposed by lifestyle. Archaea are very simple organisms that usually live in harsh and extreme environments (Gribaldo et al. 2006). We believe their extremophilic lifestyles impose constraints on their molecular make up that: (1) limit the possibility of acquiring new structures, and (2) induce a constant selective pressure to maintain a minimal structural set necessary for survival. We therefore propose that Archaea maintained a minimal set of structures while losing structures by strong reductive evolution. We note that signature As exhibit very low *f* values, suggesting these molecular designs were acquired as adaptations to new environments and lifestyles. The appearance of structures shared by only two superkingdoms was also revealing. For example, the AE boxplot's upper whisker approached nd_H = 1, implying a recent relationship between Archaea and Eukarya. Comparatively, the *nd* values for SCOP FSFs for the AE taxonomical group was nd_{FSF} = 0.85, supporting the late appearance of the interaction (Wang et al. 2007). Note that a sister relationship between Archaea and Bacteria is usually used to claim the canonical bacterial rooting of the tree of life Woese et al. (1998), but that in our studies this relationships is only supported by domain structures that are quite derived (see below).

Trees of proteomes derived from the CATH genomic census confirm the early emergence of Archaea

We previously reconstructed trees of proteomes from a genomic census of SCOP domains and made inferences about the rooting of the tree of life (Wang et al. 2007; Kim and Caetano-Anollés 2011, Kim and Caetano-Anollés 2012). We found trees of proteomes

reconstructed from ancient domain structures were rooted paraphyletically in Archaea while trees reconstructed using derived structures exhibited the canonical rooting with Bacteria emerging at their base (Figure 2.6). We also revealed how parasitic and symbiotic lifestyles can complicate phylogenetic interpretation (Wang et al. 2007; and Kim and Caetano-Anollés 2012). The proteomes of organisms that are parasitic or that establish symbiotic relationships with other organisms have frequently experienced reductive evolution, discarding enzymatic and cellular machineries in exchange for resources from their hosts. Since their inclusion can lead to incorrect phylogenetic trees, we excluded proteomes from all but 295 free-living (FL) organisms (table S2) and reconstructed most parsimonious rooted trees describing their evolution. The FL set included 41 archaeal, 189 bacterial, and 65 eukaryotic organisms. The tree of FL proteomes reconstructed from a census of H domain structures supported the trichotomy of the superkingdoms (Figure 2.6). The number of bacterial proteomes was however overrepresented in the FL-tree and could cause long-branch attraction during phylogenetic reconstruction possibly leading to incorrect deep phylogenetic relationships. Nabhan and Sarkar (2012) in a recent study have reviewd the impact of taxon sampling and long-branch attraction on a phylogenomic inference. We thus randomly sampled equal numbers of proteomes per superkingdom (a maximum of 41) and generated replicated trees of proteomes. Reconstruction of equally sampled FL proteomes improved tree resolution and bootstrap support values of deep branches. More importantly, the trees consistently showed a paraphyletic rooting in Archaea and the derived placement of monophyletic Bacteria and Eukarya (Figure 2.6). We also reconstructed trees of FL proteomes from three subsets of phylogenetic characters: ancient H structures common to all superkingdoms corresponding to the architectural diversification epoch ($nd_H < 0.176$), H structures of intermediate ancestry corresponding to the superkingdom specification epoch

(0.176 < nd_H < 0.55) and H structured that are derived and reflect the organismal diversification epoch (0.55 < nd_H). The proteome tree reconstructed from the most ancient H structures was rooted paraphyletically in Archaea, reflecting their early segregation through the minimalist strategy. Reconstructions from H structures of intermediate ancestry produced trees with three clades corresponding to the three superkingdoms that were rooted in Archaea. Finally, reconstructions from H structures that were derived yielded the canonical tree of life rooted in Bacteria. It is noteworthy that the rooting of these trees reflects the early appearance of Bacteria-specific domain structures (Figure 2.6, see trees reconstructed using most ancient, ancient and younger characters sets).

Modern Archaeo-Eukaryotic architectural sharing questions the canonical tree of life

The H structural chronology unveils a relatively recent (perhaps current) sharing of protein architectures between archaeal and eukaryal genomes that was unknown before. This inspired us to resolve the phylogenetic contribution of each structural character set in the tree of proteomes. Interestingly characters that are shared by archaeal and eukaryal genomes exhibited high retention index (RI) values (Figure 2.7), indicating that the sharing pattern did not result from annotation artifacts. The RI measures the amount of synapomorphy expected from a data set that is retained as synapomorphy on a cladogram. Boxplots of structural character sets shared by the seven taxonomical groups were also plotted (Figure 2.7). These RI boxplots are powerful enough to explain the relationships of superkingdoms in our tree of proteomes. The AE boxplot is the only one exhibiting very high RI values. In turn, bacteria-specific characters had the most dispersed RI boxplot. Hence, archaeal and eukaryotic lineages share good signal characters that are very recent and widely shared (their high f values indicate for example presence in most of archaeal and eukaryotic proteomes).

More than 30 years ago, Woese et al. (1977) reconstructed the classical (canonical) tree of life using paralogous genes. This tree delineates three domains (superkingdoms) of life – Bacteria, Archaea and Eukarya – and assumes the root branch corresponds to Bacteria. The canonical tree of life also makes the proposal that the Archaea and Eukarya are two distinct sister lineages that are derived from an exclusive common ancestor. Many archaeal components of systems involved in informational systems (e.g. translation, replication and transcription) and transmission of genetic information show a higher sequence similarity with their eukaryotic homologue than their bacterial homologue (Brown and Doolittle et al. 1997; Hartman et al. 2005). For instance, more than 30 ribosomal proteins are shared between the Archaea and Eukarya that are not present in Bacteria (Lecompte et al. 2002). Moreover Archaea and Eukarya also share a similar base excision repair system that is different than the system in bacteria (Ögrünç et al. 1998). If the signal embedded in the sequence of these RNA and protein molecules depicts history adequately, these findings explain the evolutionary link between Archaea and Eukarya and the topology of the canonical tree of life. However, the tree of proteomes reconstructed using the modern structural character set (Figure 2.6: epoch 3 or younger character set) is the only one producing the canonical tree topology that places the root branch in Bacteria. This topology mostly results from protein domain structures with very recent origin that are shared between Archaea and Eukarya. We contend that these very recent domains retain good phylogenetic signal, especially in their sequences, and will be the less affected by processes of mutation saturation. Consequently, the close evolutionary relationship of Archaea and Eukarya in trees of life derived from analyses of these sequences (Woese et al. 1977; Woese et al. 1998) can be considered an artifact of the focus on sequence.

Current trees of life built for example from sequence concatenation (e.g. Cicarelli et al. 2006; Lienau et al. 2011) include genes encoding for multidomain proteins (e.g. aminoacyl-tRNA synthetases, etc). Some of these domains are of recent origin and may fall withing the derived domain set we have analyzed. We claim that strong phylogenetic signal in the sequence of these domains likely drives the reconstructed topologies. Instead, weak phylogenetic signal embedded in the sequences of older and universal domains is swamped by the recent archaeo-eukaryotic signal that is in part responsible for the canonical tree. Our focus on CATH domain structure (not gene sequence) can dissect the differential contribution of old and recent protein domains that belong to the proteome-encoding gene repertoire. A similar focus on the deep phylogenetic signal in RNA structure has also shown the basal placement of Archaea in phylogenetic reconstructions from tRNA, RNase P RNA and 5S rRNA (Sun and Caetano-Anollés 2007; Sun and Caetano-Anollés 2008, Sun and Caetano-Anollés 2009, Sun and Caetano-Anollés 2010; Xue et al. 2003) Clearly, deep phylogenetic signal in protein and RNA structure is free from the limitations of gene sequence and associated non-vertical patterns arising from horizontal gene transfer but more importantly from domain rearrangement and can therefore reveal historical patterns without bias. Here we show the importance of considering the age heterogeneity of a biological repertoire, in this case the proteome, when making phylogenetic statements.

Chronologies of CATH architectures reveal evolutionary patterns of structural diversification

In contrast with chronologies of superfamilies and topologies (Hs and Ts), the chronology of architectures (As) is evolutionarily even more conserved (Figure 2.10). It shows that As are widely shared and are refractory to loss in genomic lineages. In fact, very few are lost

in superkingdoms (4 in Archaea, and one each in Bacteria and Eukarya). Thus, As are very old and popular in the world of organisms. The *3-layer (αβα)sandwich (3.40)* is the most abundant and ancient of all proteins. The *orthogonal bundle (1.10)* and the *α/β–complex (3.90)* are equally abundant and are the second and third most ancient architectures. Remarkably, the phylogenomic tree of As shows that comparatively simpler shape structural designs are more favored than complex designs and in general are more ancient, appearing at the base of the tree. For example, the most ancient *3.40* and *1.10 architectures* involve simple arrangements of secondary structure while more recent shape designs are more complex (Figure 2.9). As the time progresses the complexity in architectural make up of structural designs also increases, a general trend observed in the tree of As. The few As that are lost in superkingdoms are quite complex and as expected their appearance is quite derived. The first loss occurred in Eukarya (nd_A = 0.76) with the very complex *Clam* architecture, and then in Archaea and Bacteria. We note that Archaea loses four As quite late and in a row, showing that the pervasive reductive trends of Archaea described above extend almost to the present. This also reflects the conservative nature of extremophilic Archaea, which are not in need of modern structural designs. Bacteria loses the most recent A structural design, *Box* (*2.80*), at nd_A = 1, which is shared by both archaeal and eukaryal genomes. *Box* is involved in nucleotide excision repair, a molecular function that has a unique place in cellular defense because of its wide substrate range and its ability to virtually remove all base lesions from a genome. Ögrünç et al. (1998) reported a similar base excision repair system used in Archaea and Eukarya and argued that a different set of proteins are employed by the bacterial nucleotide repair system. Interestingly, the *f* index for *Box* in Archaea (*f* = 1) and Eukarya (*f* = 0.96) again indicates a recent sharing of structural designs between archaeal and eukaryal organisms. Architectures are the second highest level of structural abstraction in CATH, and

because of their high conservation it is difficult to clearly delimit the three epochs of the protein world. In contrast, our results indicate CATH H and SCOP FSF are the most suitable levels to uncover the evolution of domain structures in genomes. These levels of abstraction are structurally and evolutionarily conserved. They preserve deep phylogenetic signatures and variable enough to dissect evolutionary history of proteomes and molecular functions.

CATH architectures become more complex in evolution

The structural make up of the most ancient *3-layer (αβα) sandwich (3.40)* architecture represents the central theme of the most ancient SCOP FFs (Caetano-Anollés et al. 2012). These structures consist of repeating α–β–α supersecondary units, such that the outer layer of the structure is composed of helices packing against a central core of parallel β–sheets. Many enzymes, including most of those involved in glycolysis, are α/β layered proteins and are cytosolic (Branden and Tooze 1999). These α/β structures harbor repetitions of the α–β–α arrangement (e.g., the α–β–α–β–α sequence). The β-strands are parallel and hydrogen bonded to each other, while the α-helices are all parallel to each other but are antiparallel to the strands. Thus the helices pack against the sheet forming a sandwich like structure. We note that the β–**α–β–α**–β (αβα) subunit, often present in nucleotide-binding proteins, represents the *Rossmann* structural motif found in proteins that bind nucleotides, especially the cofactor NAD(H) (Rao and Rossmann 1973).

The *orthogonal bundle (1.10)* and *α/β-complex (3.90)* appear immediately after the *3-layer (αβα) sandwich (3.40)* design. The *orthogonal bundle* consists of a 3-4 α–helix bundle and is found in a number of different proteins, most of which associate with membranes. Due to physical constraints imposed by the lipid bilayer of membranes the list of possible membrane

protein structures is limited to either bundles (Rees et al. 1989; Wallin et al. 1997) or barrels (Weiss et al. 1991; Wimley et al. 2003). In many cases the α–helices are part of a single polypeptide chain and are connected to each other by three loops. In the 4-helix bundle proteins the interfaces between the helices consist mostly of hydrophobic residues while polar side chains on the exposed surfaces interact with the aqueous environment. A number of cytokines consist of 4-helix bundles, such as *interleukin-2*, *interleukin-4*, *human growth hormones*, and the *granulocyte-macrophage colony-stimulating factor (GM-CSF)* (Branden and Tooze 1999) and DNA binding proteins (e.g., transcription factors, repressors proteins) (Stargell et al. 2001). CATH has grouped the complex shaped structures into the 'complex' bin, until alternative assignment methods are developed. The *α/β-complex* architecture groups together all those designs that include significant α and β secondary structural elements in a mixed fashion. Examples of *α/β-complex* proteins include bacterial and mammalian *pancreatic ribonucleases* (Scheraga et al. 2001), *Zn metallo-proteases* and *DNA topoisomerases* (Giangreco et al. 2011). Two kinds of *barrel* structures are the most ancient and abundant in the protein world, the *α/β–barrel (3.20)* and the *β–barrel (2.40)* (Orengo et al. 1997), and both appeared on the same time (nd_A = 0.13). The *α/β–barrel* is composed of eight α–helices and parallel β–strands that alternate along the peptide backbone. The *α/β–TIM barrel* is the most prominent example of *α/β–barrel* and is widely present in enzymes of central metabolism (Lee and Herman 2011). A *β–barrel* is a large β–sheet that twists and coils to form a closed structure in which the first strand is hydrogen bonded to the last. β–strands in *β–barrels* are typically arranged in an antiparallel fashion. Barrel structures are commonly found in porins and other proteins that span cell membranes and in proteins that bind hydrophobic ligands in the barrel center, such as *lipocalins* (Campanacci et al. 2004). The *roll* is a complex nonlocal structure in which 3-4 pairs

of antiparallel β–sheets, only one of which is adjacent in sequence, are 'wrapped' in 3D space to form a barrel shape (Andreeva et al. 2010). Rolls appear for the first time at nd_A = 0.3.

A number of distinct and more complex architectures appear later on in the chronology, including *solenoids, horseshoes, prisms, propellers and trefoils*. *Solenoid* proteins, with their arrays of repeating motifs, tend to have elongated structures that contrast with the majority of globular proteins whose polypeptide chains follow more complex trajectories (Forwood et al. 2010). These are constructed from tandem structural repeats arranged in superhelical fashion, a feature that is important for many cellular processes (Kobe and Kajava 2000). Solenoid proteins constructed from HEAT repeats (Groves et al. 1999) and *armadillo* repeats (Kobe et al. 1999; Peifer et al. 1994) constitute the principal transport receptors. A key structural property that differentiates solenoid proteins from other structured proteins is the lack of contacts between distal regions of protein sequence (sequence-distal contacts). For this reason, solenoid proteins are often more flexible than other structured proteins and this flexibility is an important feature of their specific functions (Forwood et al. 2010). Solenoid structure appears for the first time at nd_A = 0.46. The *α–horseshoe* protein appears at nd_A = 0.4, is a super helical structure made up of a number of 3 α–helical orthogonal bundle repeats. The α–β horseshoe appeared at nd_A = 0.56, consists of several α/β–repeating units (Kobe and Deisenhofer et al. 1993). The structure of the *ribonuclease Inhibitor*, a cytosolic protein that binds strongly to any *ribonuclease* that may leak into the cytosol, takes the concept of the repeating α/β unit to the extreme (Kobe and Deisenhofer et al. 1993). The structure is made of a 17-stranded parallel β–sheet curved into an open horseshoe shape, with 16 α–helices packed against the outer surface. *Prisms* are similar to *solenoids* in geometry but completely different in connectivity. A more self-contained β–sheet forms each face of a triangular prism. They appear late at nd_A = 0.86. The *trefoils* consist of an

unusual β sheet formed by six β hairpins arranged with three fold symmetry into 'Y' like structures (Taylor and Aszodi 2005) and are also quite derived (nd_A = 1).

To obtain a detailed view of architectural discovery and usage over time, we mapped the appearance of H and T structures harboring individual A designs, plotting nd_H and nd_T values for Hs and Ts belonging to each of the 38 known As (Figure 2.12). We also grouped the As into 10 major structural designs: *sandwiches*, *bundles*, *barrels*, *prisms*, *horseshoes*, *rolls*, *solenoids*, *propellers*, *complexes* and *other* (a category that contain structural designs that could not be clearly grouped into the main categories) (Table 1 and Figure 2.11). We found that most *sandwiches*, *bundles*, *barrels*, *complexes* and *rolls* have high *f* values (*f* ~ 1) and rather simple structural designs (Figure 2.11). In turn, structural designs such as *propellers*, *horseshoes*, *solenoids (2 Solenoid, 2.150), prisms, trefoil* and *box*, have low *f* values (*f* = 0.85-0.10) and are very complex. These complex designs appeared late in evolution and as expected are sparsely distributed in the world of organisms.

Models of evolution of CATH and SCOP domain structures are congruent

The most ancient and popular architecture, the *3-layer (αβα) sandwich* (*3.40*), harbors the most ancient and abundant topology, the *Rossmann fold* (*3.40.50*) and the most ancient and abundant superfamily, the *P-loop containing nucleotide triphosphate hydrolases* (*3.40.50.300*). Despite differences of topology and ranking within databases (Chu et al. 2005), this H structure of CATH is analogous to the "*P-loop containing nucleotide triphosphate hydrolase*" FSF (c.37.1) of SCOP (Casba et al. 2009), since both have Rossmann fold topology and also agree on their keyword definitions. A careful analysis of CATH and SCOP structures phylogenies show that the ancient domains structures at T (3.40.50) and H (3.40.50.300) levels are in global

agreement with timelines of F (c.37) and FSF (c.37.1) structures (Wang et al. 2007). Despite differences in domain definitions of tertiary structure in CATH and SCOP, the remarkable conservation of evolutionary signal indicates both classification systems effectively preserve evolutionary information in protein structure and uncover global patterns of origin and diversification that are for the most part congruent.

2.4 CONCLUSIONS

In this study we use a radical approach to analyze the evolution of protein fold structure and proteomes in the tripartite world of organisms. Instead of generating trees of life from protein sequence with standard methods, we use a genomic structural census and robust cladistics methods to build trees of domain structures and proteomes. Structural phylogenies describing the evolution of CATH domains at A, T and H levels of structural abstraction revealed patterns of reductive evolution and three epochs in the evolution of the protein world that were previously proposed (Wang et al. 2007). Structural diversification patterns match those observed in the analysis of SCOP domain structures (Wang et al. 2007; Yang and Song 2009; Kim and Caetano-Anollés 2012). Reconstruction of phylogenomic trees of proteomes describing the evolution of lineages confirms Archaea is the most ancient superkingdom. Five major findings summarize novel results and take advantage of the ability of CATH to better describe topological features of protein structure:

1. Structural designs that are architecturally simpler are ancient and highly abundant in the extant world of proteins and organisms. We find the *3-layer (αβα) sandwich* cytosolic architecture and the *orthogonal bundle* that often associates with membranes are the most ancient and are preferentially involved in metabolic activities. The origin of proteins thus

lies at the interface of primordial membranes and cytoplasm. Bundles and barrels that populate membranes soon follow. Metabolic and membrane proteins thus appear crucial for the early biochemistry of primordial cells (Caetano-Anollés et al. 2012).

2. Structural designs that are architecturally complex, such as *prisms, propellers, 2-solenoid, super-roll, clam, trefoil and box* are derived and less favored in the world of organisms. These designs are generally specific to groups of organisms and have been probably adopted for specialized functions.
3. Although CATH and SCOP differ significantly in their protein domain definitions and in the hierarchical partitioning of fold space, we find that both protein structural classification systems classify a protein on very similar theoretical grounds by taking into account their structural, functional and evolutionary roles. Remarkably, CATH and SCOP structures harbor similar phylogenetic signatures and reveal patterns of origin and diversification that are congruent.
4. Structural chronologies provide evidence that Archaea established the first organismal divide by losing a substantial number of domain structures early in evolution. We speculate this reductive evolutionary process reflects the environmental pressure of an ancient extremophilic lifestyle that forced maintenance of a minimal domain repertoire.
5. Structural chronologies uncover a recent trend of sharing of domain structures between Archaea and Eukarya that continues to the present and involves complex architectures such as the *Box* (2.80) design that is involved in nucleic acid repair.
6. Finally, we also speculate that this modern archaeo-eukaryotic architectural sharing pattern is the most probable reason for the bacterial rooting of the canonical tree of life usually derived from changes in sequence. In contrast to structure, sequence evolution is more

dynamic and prone to phylogenetic signal loss. It is therefore likely that most useful phylogenetic signal in these sequence studies is drawn from structures that have been developed quite recently in evolution.

Our trees of domain structures define timelines that trace back the history of discovery, diversification and distribution of protein structural designs. Our finding that protein architectures tend to become more complex in evolution is very significant. In a previous study, analysis of β–barrel structures revealed that the curl and stagger and complexity of the connectivity of supersecondary structures increases in evolution (Caetano-Anollés and Caetano-Anollés 2003). The very early appearance of multilayered sandwich structures is also compatible with the finding that the most ancestral folds share a common architecture of interleaved β–sheets and α–helices (Caetano-Anollés and Caetano-Anollés 2003). An even more recent study shows that 36 out of the 54 most ancient FFs harbor α/β/α–layered sandwich structures (Caetano-Anollés et al. 2012). The very early appearance of the P-loop hydrolase motif in the first FF, the ABC transporters, was associated with a built-in lateral bundle, which resembles the transmembrane domains of transporter proteins. This suggests that first proteins contained sandwich and bundle structures and were associated with the membranes of primordial cells. Remarkably, P-loop hydrolase folds and bundles make up important membrane complexes, such as ion channels and transporters. Their very early origin highlights a crucial links between the origin of proteins and the origin of cells.

Figures

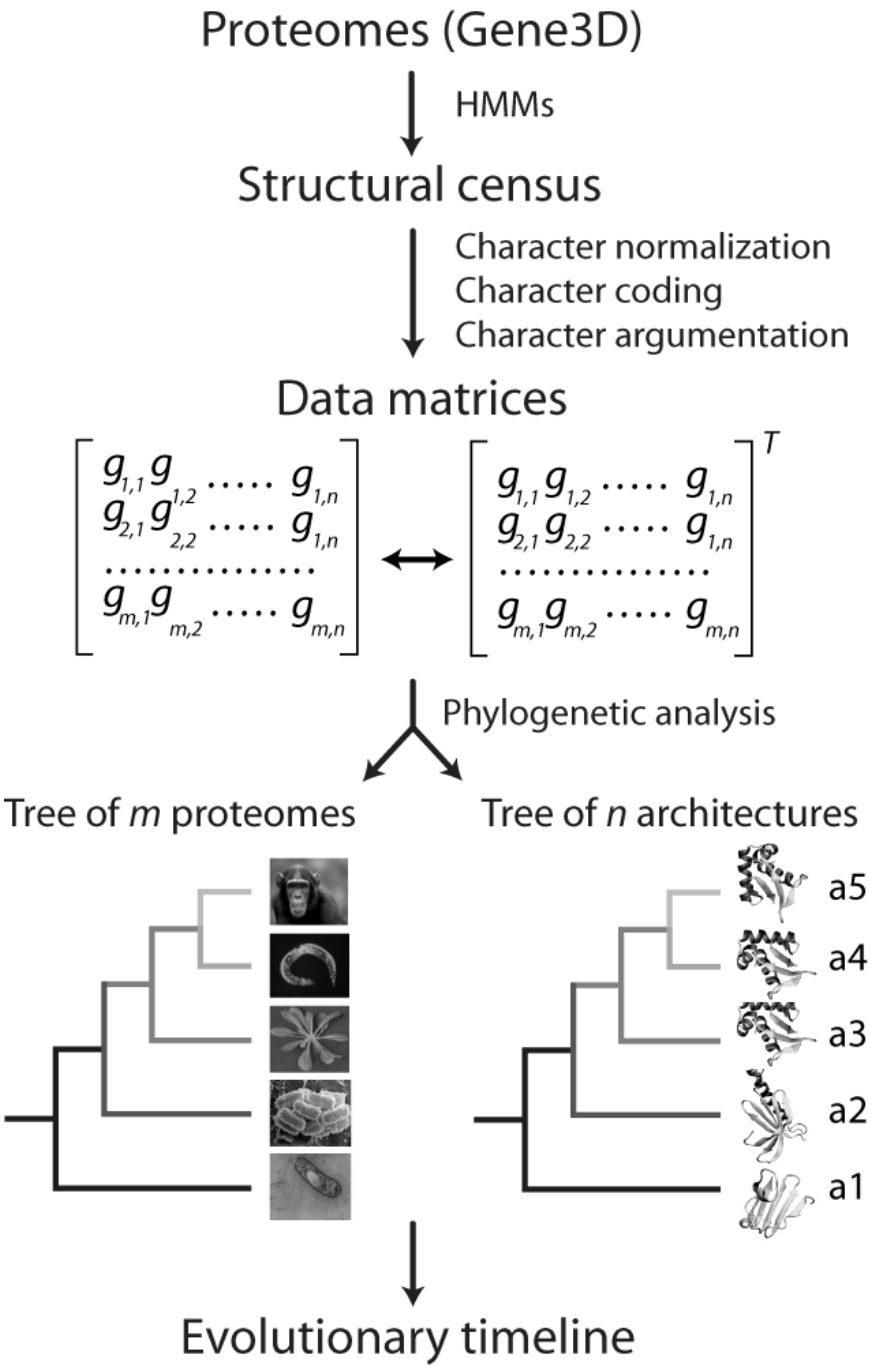

Figure 2.1 A flowchart of methodology adopted for reconstructing phylogenies, for protein architectures, and for tree of proteomes using protein domains census data.

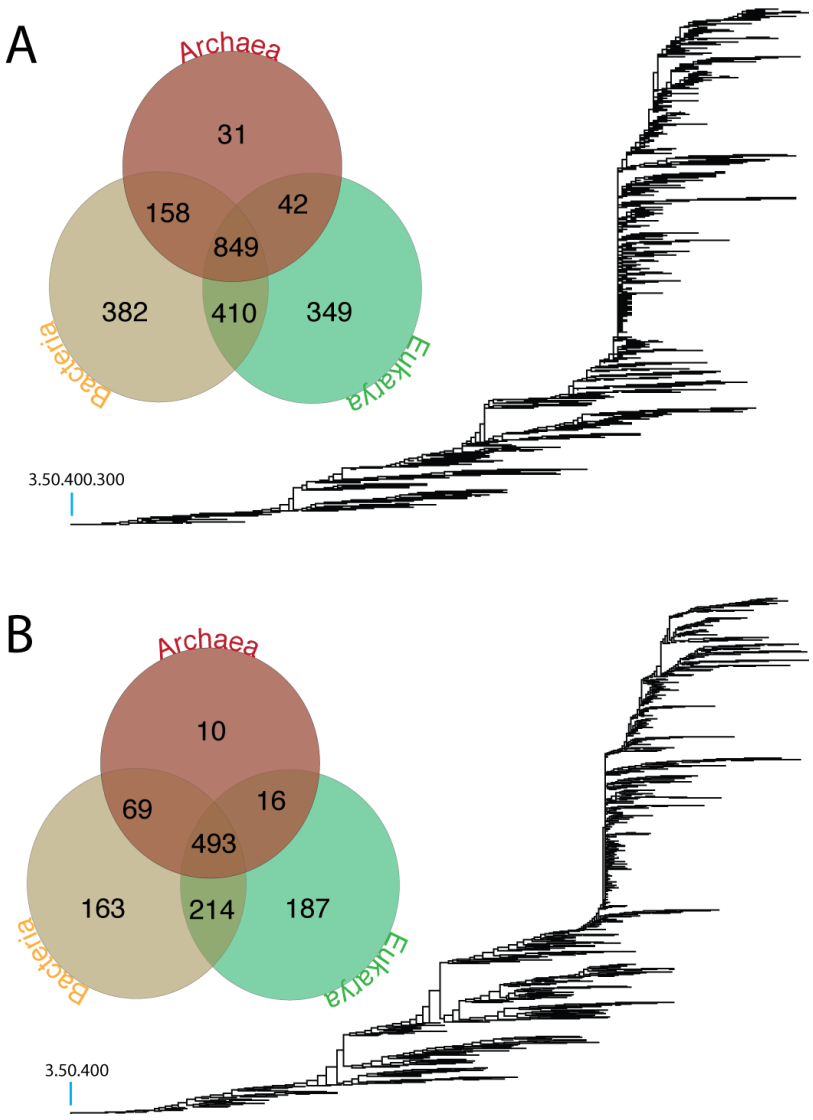

Figure 2.2 (A) Phylogenomic tree of H domain structures reconstructed from a genomic census of 2,221 Hs in 492 proteomes, where all 492 characters were parsimoniously informative. Terminal leaves are not labeled because they would not be legible. The Venn diagram shows the diversity of H in the three superkingdoms. (B) Phylogenomic tree of T domain structures reconstructed from a genomic census of 1,152 Ts in 492 proteomes, where all 492 characters were parsimoniously informative. Terminal leaves are not labeled because they would not be legible. The Venn diagram shows the diversity of T in the three superkingdoms.

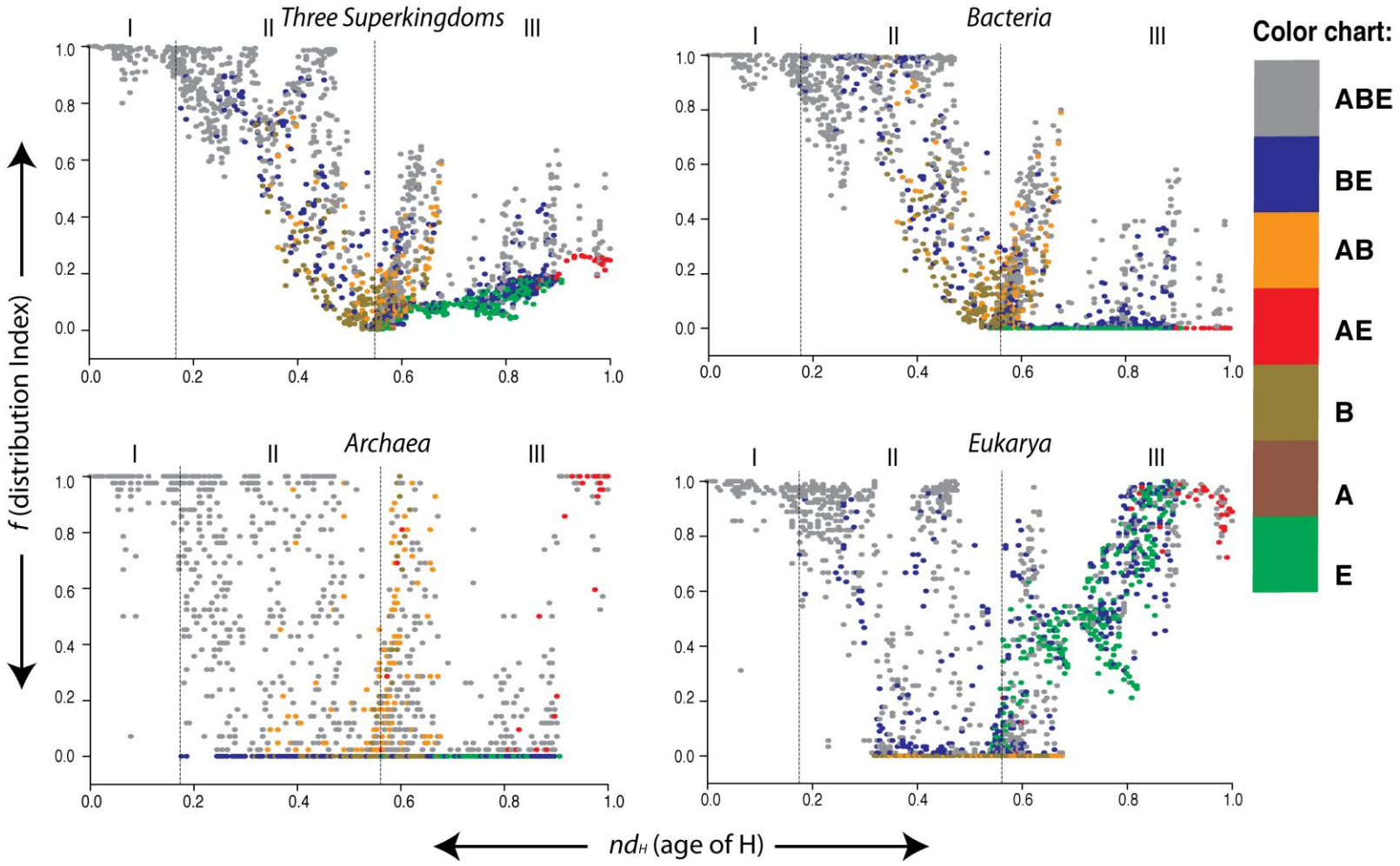

Figure 2.3 Three phases (also known as epochs) in the evolutionary timeline of appearance of H in all three superkingdoms (top-left), and in Archaea, Bacteria, and Eukarya. Individual plots show the relationship of f (distribution Index) and nd values (age of H).

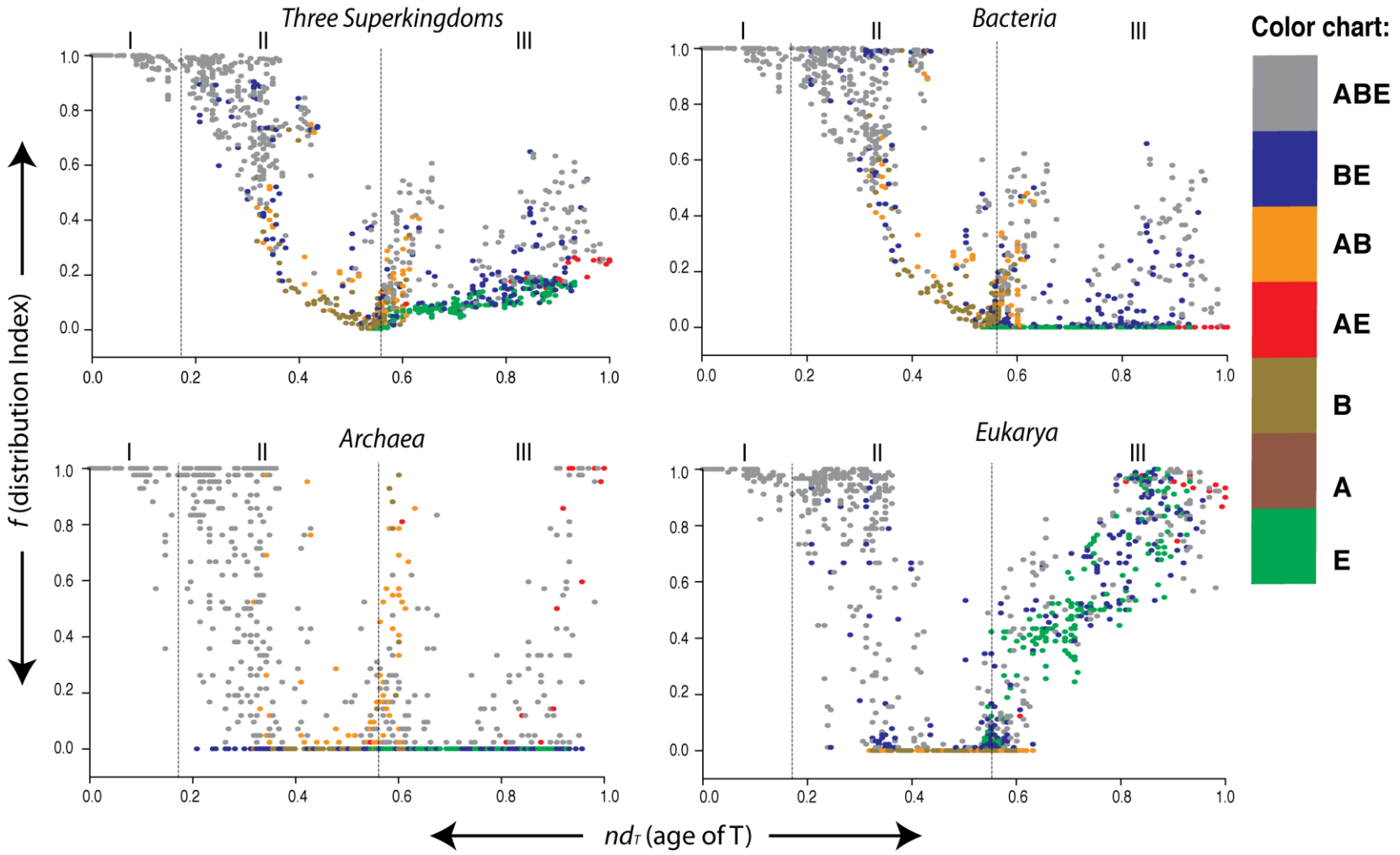

Figure 2.4 Three phases (also known as epochs) in the evolutionary timeline of appearance of T in all three superkingdoms (top-left), and in Archaea, Bacteria, and Eukarya. Individual plots show the relationship of f (distribution Index) and nd values (age of T).

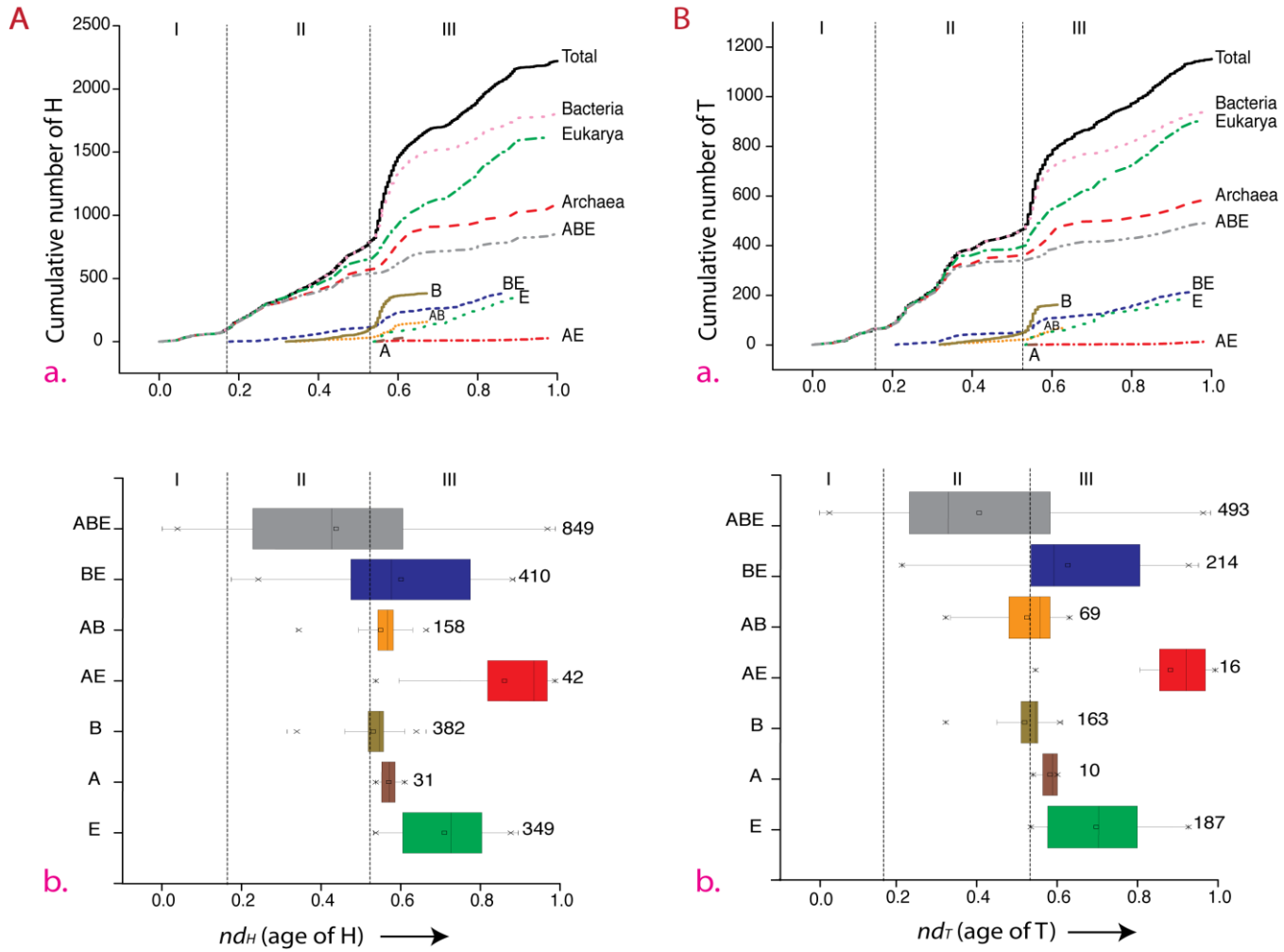

Figure 2.5 In (A.a) and (B.a) cumulative frequency distribution of H and T were plotted along the timelines of H and T domain structures respectively. (A.b) and (B.b) describes the seven boxplots indicate nd ranges for taxonomic groups of H and T that are unique to individual superkingdom (A, B, E) or shared by two (AB, BE, AE) or all (ABE) superkingdoms.

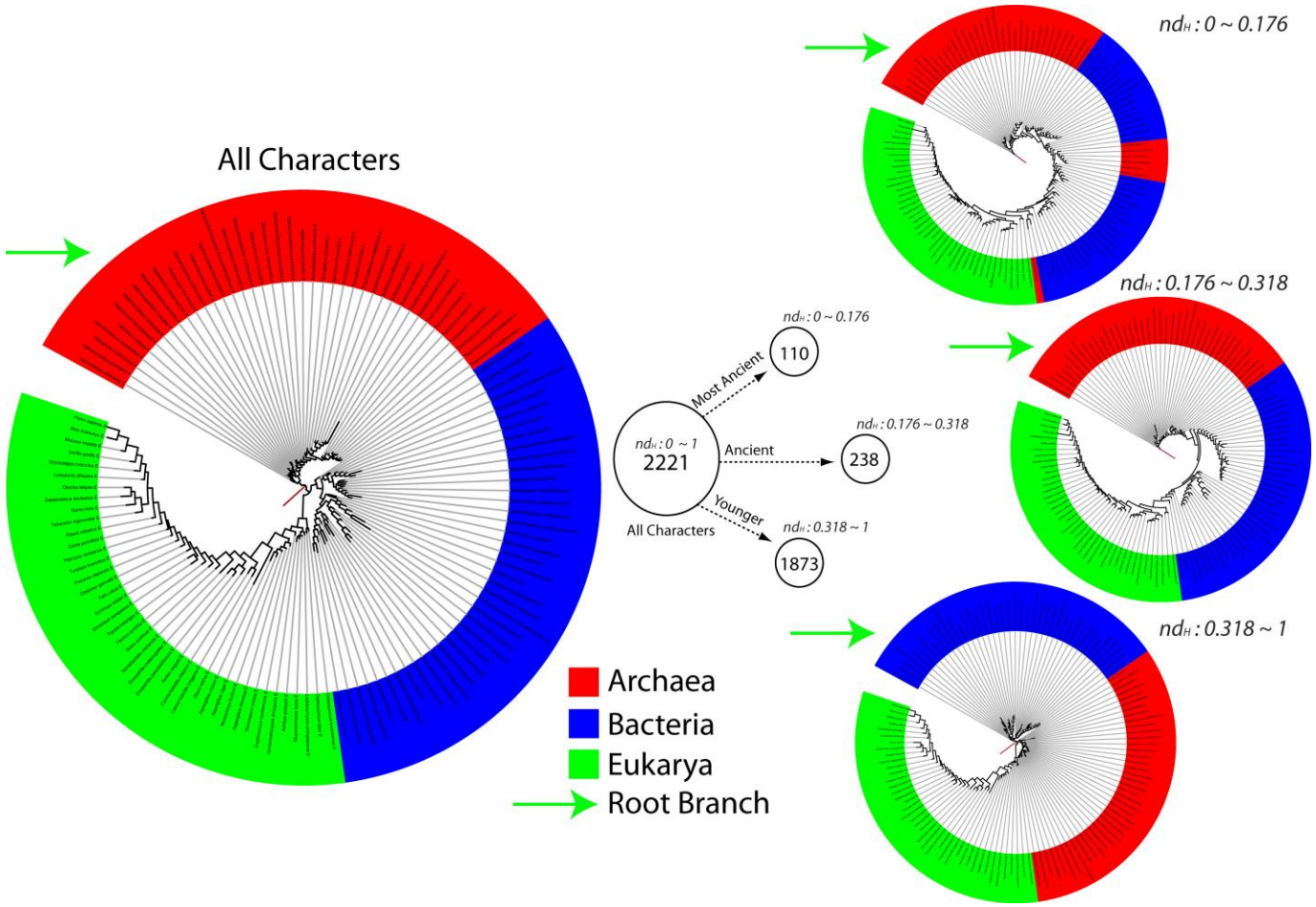

Figure 2.6 A phylogenomic tree of proteomes generated from the equally sampled dataset of FL proteomes. The circular cladogram of the most parsimonious rooted tree describes the evolution of 123 equally sampled proteomes and was generated from genomic abundances of 2221 Hs. Terminal nodes of Archaea (A: 41 proteomes), Bacteria (B: 41), and Eukarya (E: 41) were labeled in red, blue, and green, respectively. Also the total character set was divided into three independent character sets e.g. Most Ancient (nd_H 0 ~ 0.176), Ancient (nd_H 0.176 ~ 0.318) and Younger (nd_H 0.318 ~ 1) characters set. These character sets resulted in three trees of proteomes that reflected the behavior of the tree over different character sets.

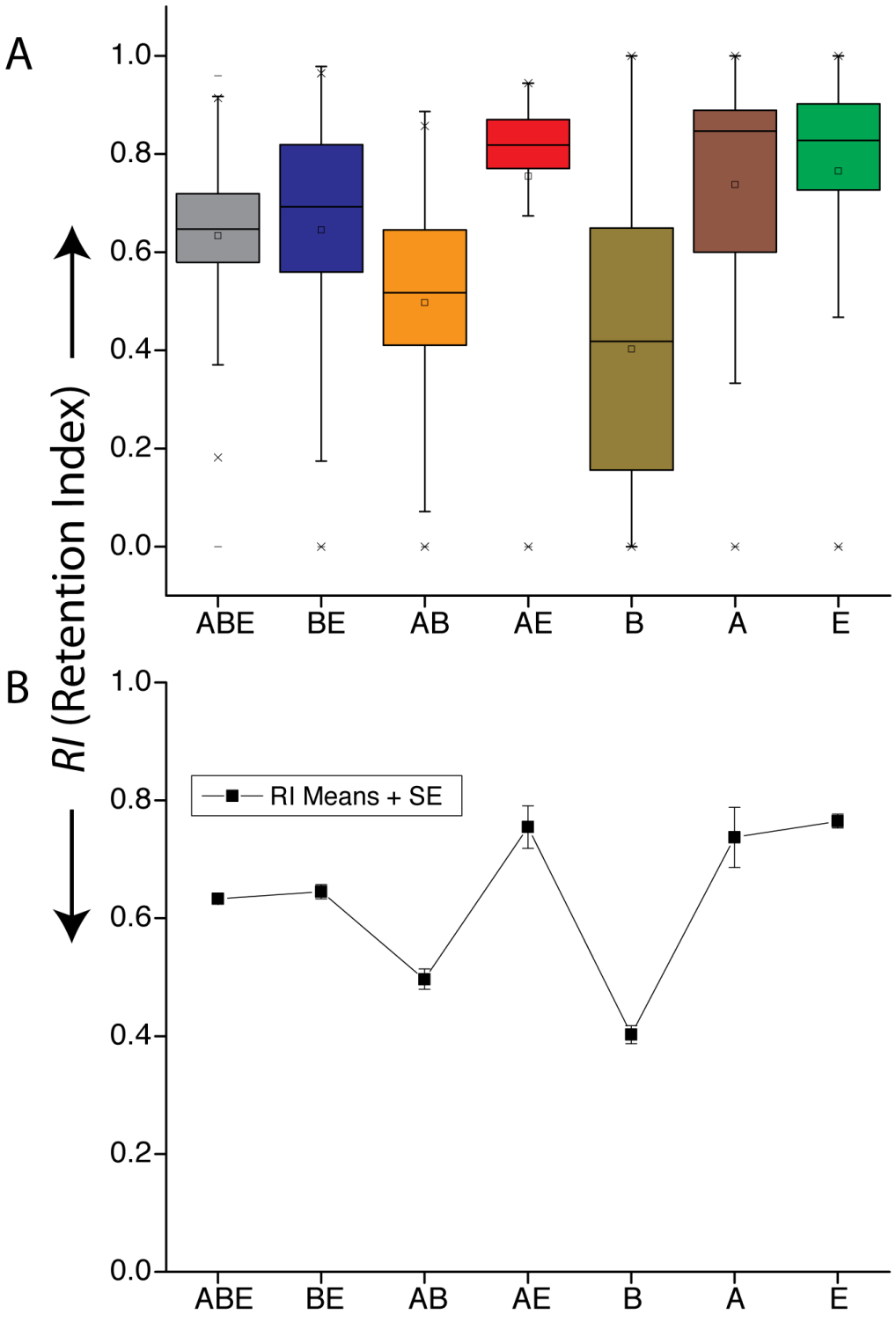

Figure 2.7 (A) Boxplots for retention index (RI) values of characters specific to seven taxonomical groups. (B) Mean RI for each taxonomical group was plotted with its standard error.

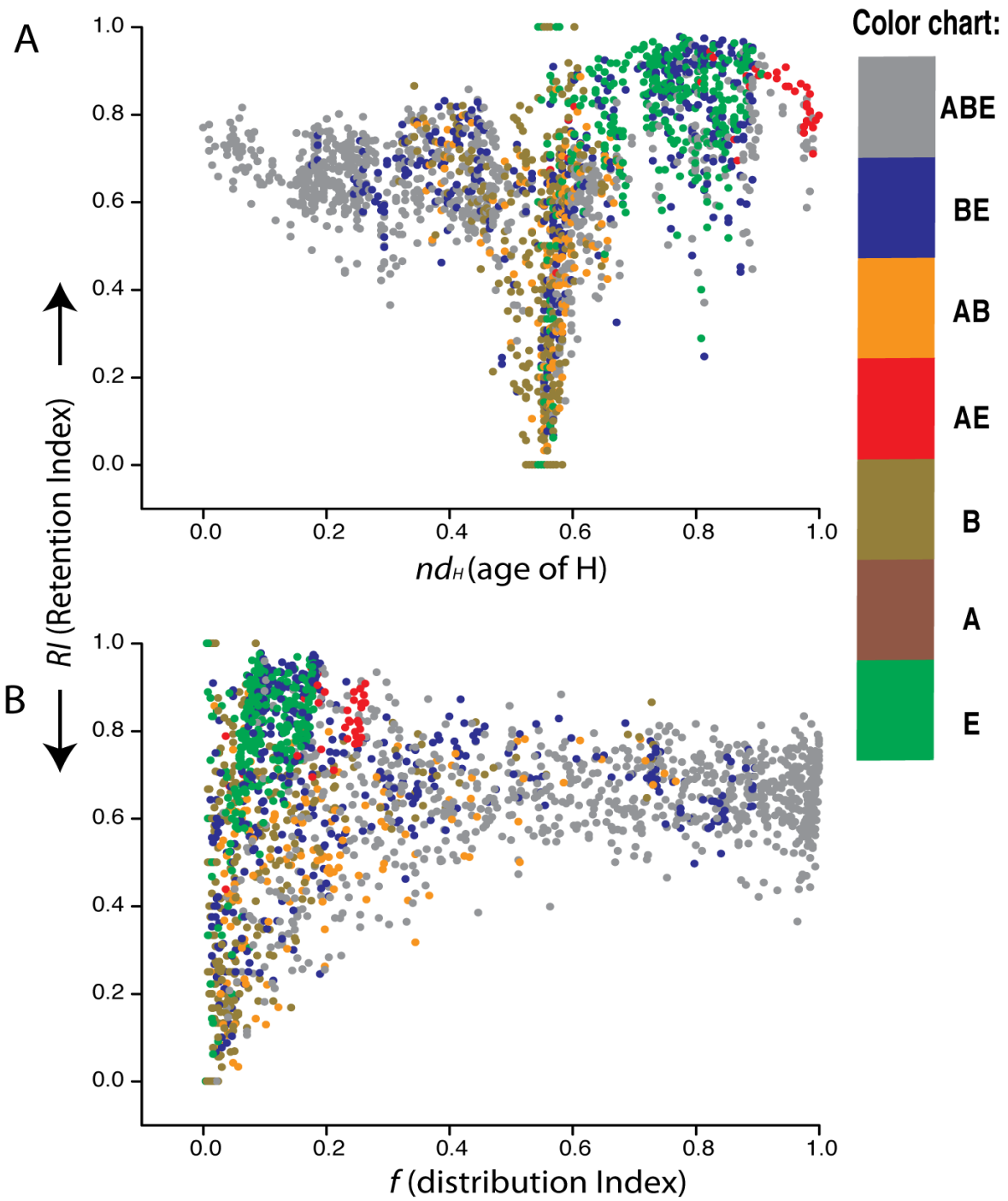

Figure 2.8 The extent of synapomorphy exhibited by phylogenomic characters (H) in the trees of proteomes. (A) RI is plotted against the age (ndH) of each character, colored according to its specific taxonomical group. (B) RI is plotted against the f distribution index of each, same coloring scheme were used as of (A).

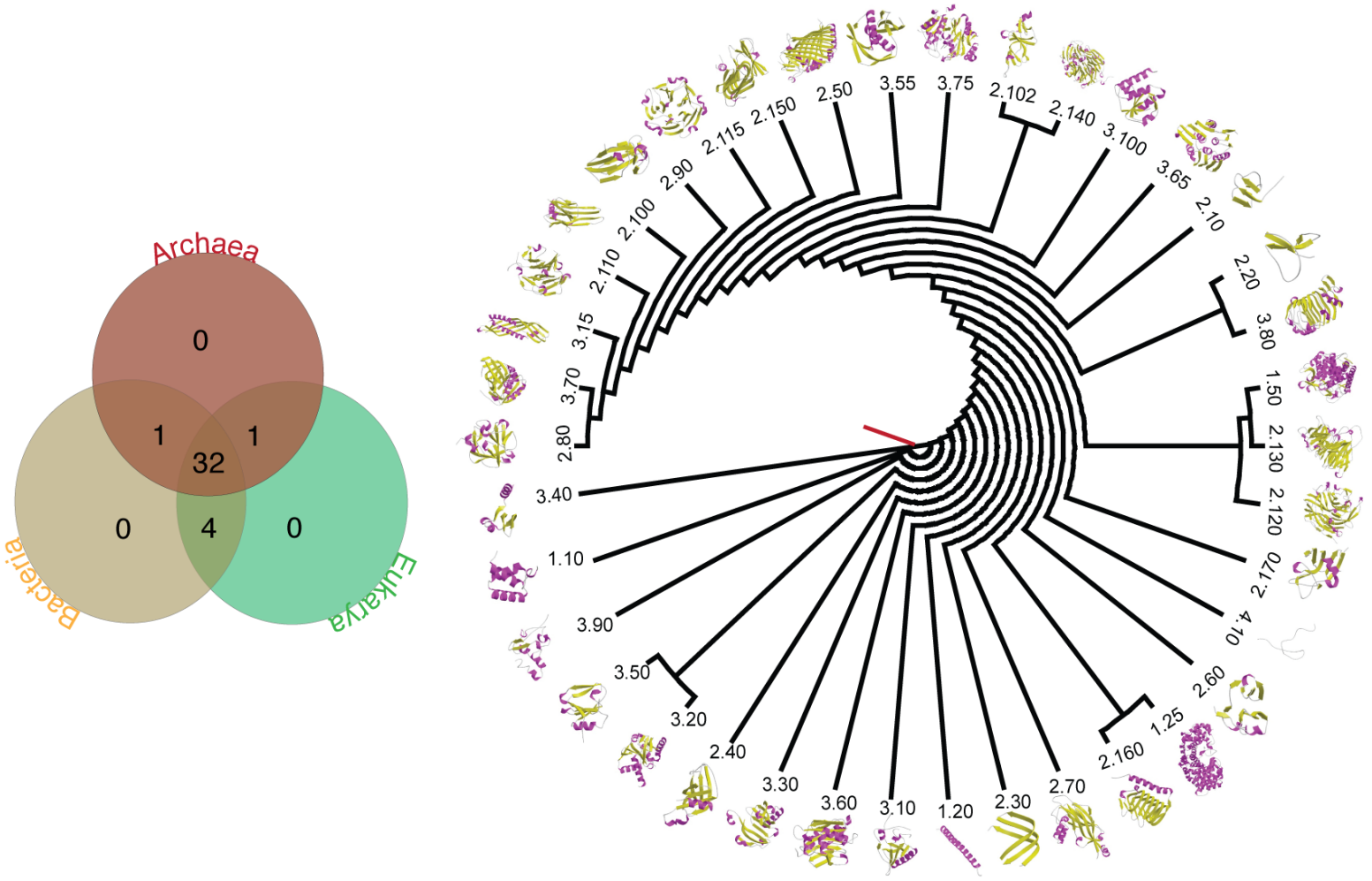

Figure 2.9 An evolutionary tree of CATH A level was plotted into circular tree topology and also A cartoon representation were mapped on each A CATH id. The Venn diagram shows the diversity of A in the three superkingdoms.

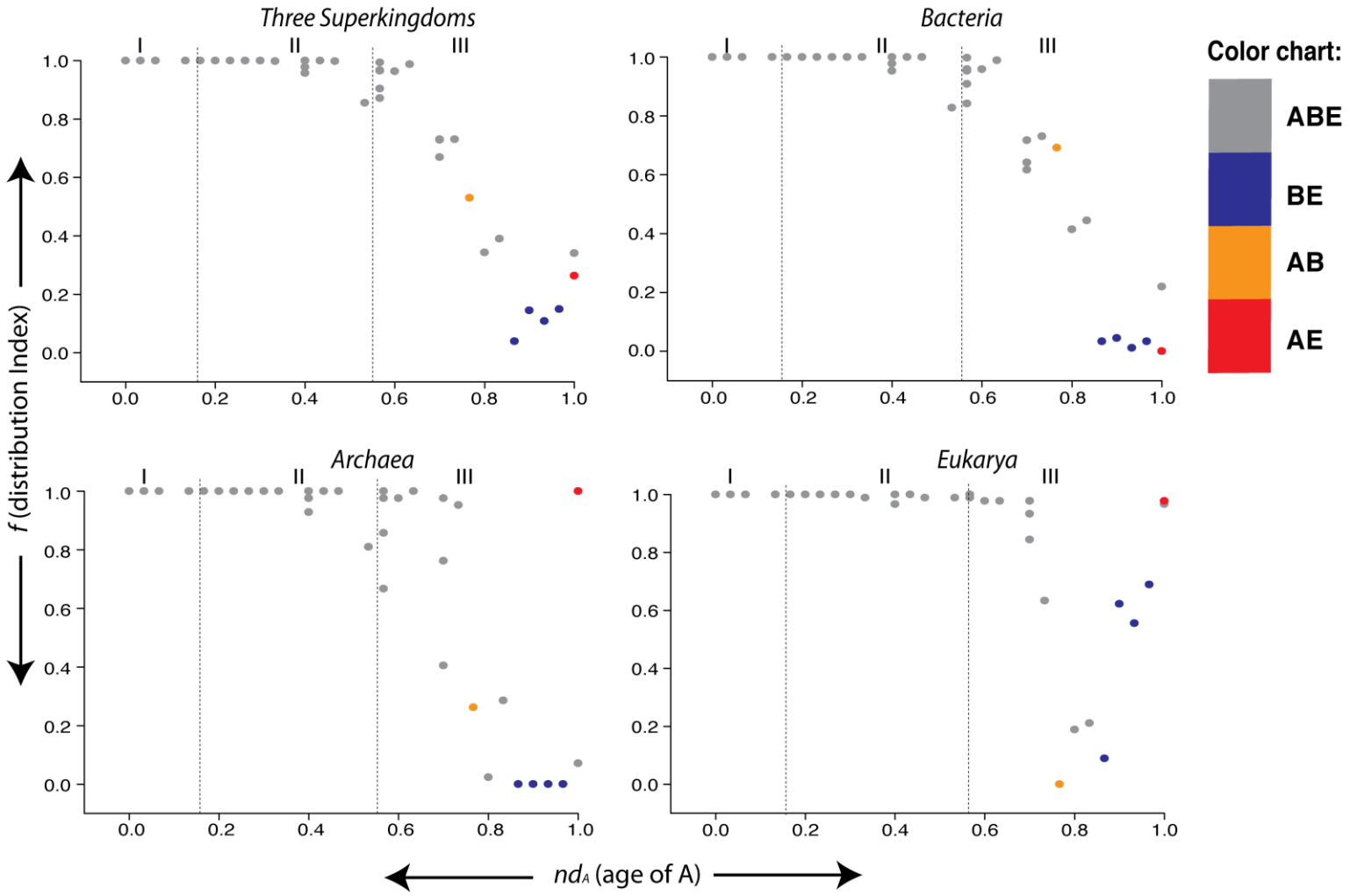

Figure 2.10 Three phases (also known as epochs) in the evolutionary timeline of appearance of A in all three superkingdoms (top-left), and in Archaea, Bacteria, and Eukarya. Individual plots show the relationship of *f* (distribution Index) and nd values (age of A).

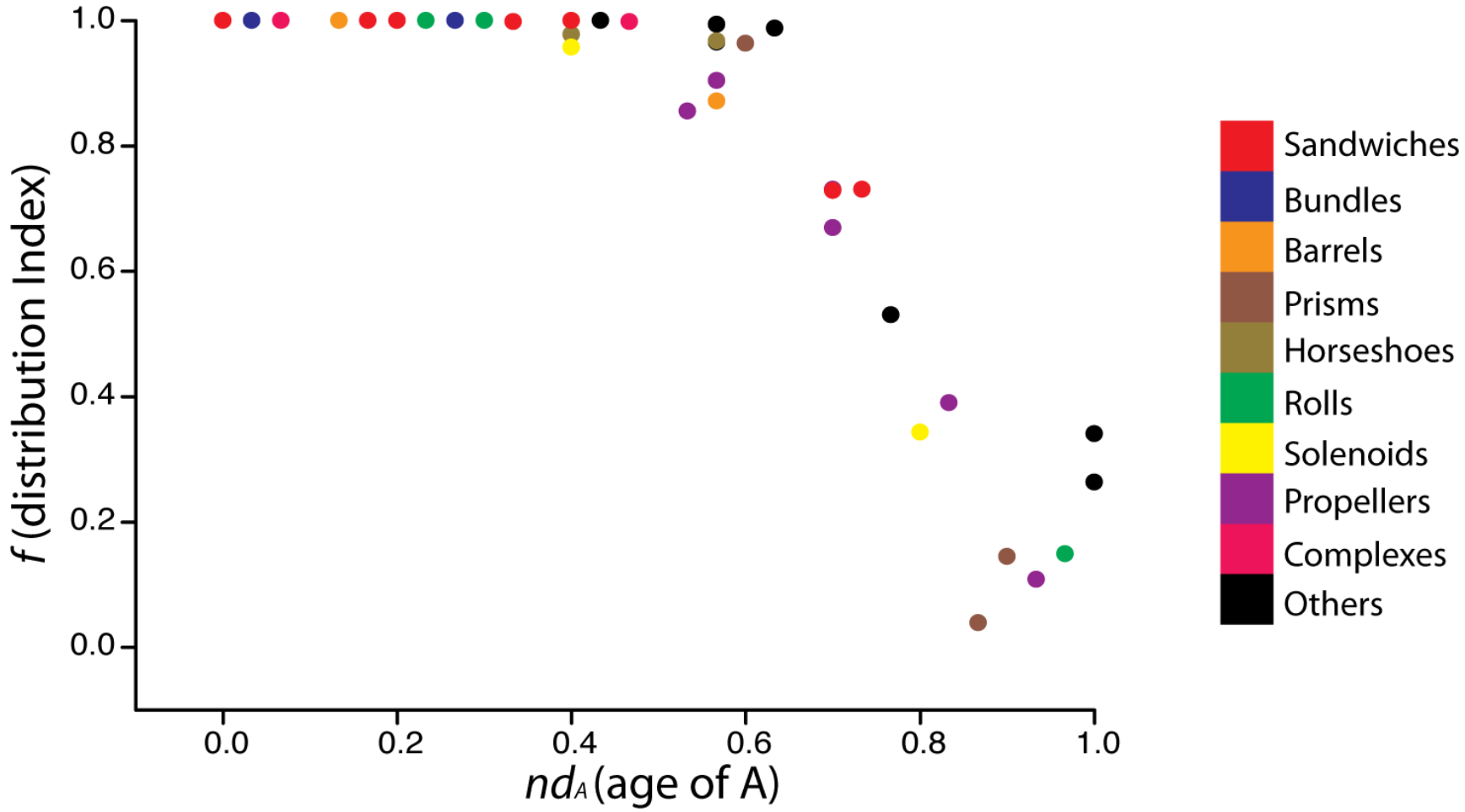

Figure 2.11 As shown in table 1 we grouped the 38 A in 10 larger sets of general structural designs. A were plotted against its age (nd_A) and *f* distribution Index, whereas each A was colored according to their general structural design group.

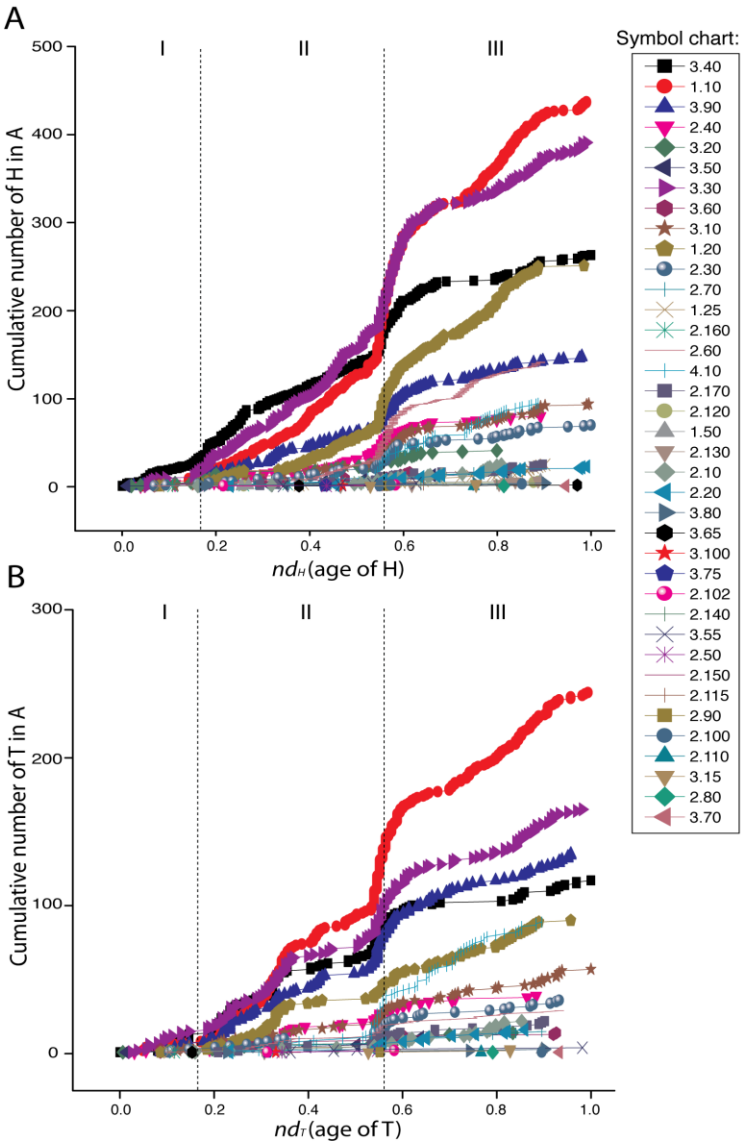

Figure 2.12 (A) & (B) Cumulative frequency distribution of T and H belonging to a particular A, along the timeline of A domain structures. Both plots depict the evolution of appearance (or recruitment) of T and H domain structures in each structural design over the timeline. Many interesting findings can be deduced from (A) like the oldest architecture 3-layer (αβα) sandwich (3.40) did not diversified itself like the orthogonal bundle (1.10), 2-Layer Sandwich (3.30) and α–β complex (3.90) did. However this pattern doesn't persist when we go one level down in (B), where 4-Layer sandwich H accumulation curve beats α–β complex as is in (B). These also indicate that instead of diversifying, 3.40 increased its popularity among the world of organisms.

Tables

Table 2.1 Generalized grouping of CATH A level into 10 general categories. It describes the age (nd_A), f distribution index and generalized structural design grouping for 38 CATH A found in our dataset with two letter CATH code and keyword description.

Index	*CATH A ID*	*CATH A Description*	nd_A	f_A	*General Groups*
1	3.40	3-Layer (αβα) Sandwich	0	1	Sandwich
2	1.10	Orthogonal Bundle	0.03	1	Bundle
3	3.90	α–β Complex	0.06	1	Complex
4	2.40	β Barrel	0.13	1	Barrel
5	3.20	α–β Barrel	0.13	1	Barrel
6	3.50	3-Layer (ββα) Sandwich	0.13	1	Sandwich
7	3.30	2-Layer Sandwich	0.16	1	Sandwich
8	3.60	Up-down Bundle	0.2	1	Bundle
9	3.10	Roll	0.23	1	Roll
10	1.20	4-Layer Sandwich	0.26	1	Sandwich
11	2.30	Roll	0.3	1	Roll
12	2.70	3 Solenoid	0.33	0.99	Solenoid
13	1.25	Distorted Sandwich	0.4	0.97	Sandwich
14	2.160	Sandwich	0.4	0.95	Sandwich
15	2.60	α Horseshoe	0.4	1	Horseshoe
16	4.10	Irregular	0.43	1	Others
17	2.170	β Complex	0.46	0.99	Complex
18	2.120	6 Propellor	0.53	0.85	Propellor
19	1.50	α/β barrel	0.56	0.87	Barrel
20	2.130	Ribbon	0.56	0.90	Others
21	2.10	Single Sheet	0.56	0.99	Others
22	2.20	7 Propellor	0.56	0.96	Propellor
23	3.80	α–β Horseshoe	0.56	0.96	Horseshoe
24	3.65	α–β prism	0.6	0.96	Prism
25	3.100	Ribosomal Protein L15	0.63	0.98	Others
26	3.75	3-layer Sandwich	0.7	0.73	Sandwich
27	2.102	8 Propellor	0.7	0.72	Propellor
28	2.140	5-stranded Propeller	0.7	0.66	Propellor
29	3.55	3-Layer (βαβ) Sandwich	0.73	0.73	Sandwich
30	2.50	Clam	0.76	0.53	Others
31	2.150	2 Solenoid	0.8	0.34	Solenoid
32	2.115	5 Propellor	0.83	0.38	Propellor
33	2.90	Orthogonal Prism	0.86	0.03	Prism
34	2.100	Aligned Prism	0.9	0.14	Prism
35	2.110	4 Propellor	0.93	0.10	Propellor
36	3.15	Super Roll	0.96	0.14	Roll
37	2.80	Trefoil	1	0.34	Others
38	3.70	Box	1	0.26	Others

CHAPTER 3

REFERENCES

Ahmed Ragab Nabhan and Indra Neil Sarkar (2012) The impact of taxon sampling on phylogenetic inference: a review of two decades of controversy. Briefings in Bioinformatics, Volume 13, Pages 122-134.

Andreeva, A. and G. Murzin, AG. (2010) Structural classification of proteins and structural genomics: new insights into protein folding and evolution. Acta crystallographica. Section F 66(10), pp. 1190-1197.

Andreeva, A. and Murzin, AG. (2006) Evolution of protein fold in the presence of functional contraints. Current Opinion in Structural Biology, 16 (3), pp. 399-408.

Bernstein, FC. Koetzle, TF. Williams, GJ. Meyer, EF. Brice, MD. Rodgers, JR. Kennard, O. Shimanouchi, T. Tasumi, M. (1977) The protein data bank: a computer-based archival file for macromolecular structures. Journal of Molecular Biology 112 (3), pp. 535-542.

Branden, C. Tooze, J. (1999). Introduction to Protein Structure 2nd ed. Garland Publishing: New York, NY.

Brown, JR. and Doolittle, WF. (1997) Archaea and the prokaryote-to-eukaryote transition. Microbiology Molecular Biology Review 61(4) , pp. 456-502.

Bussemaker, HJ. Thirumalai, D. and Bhattacharjee, JK (1997) Thermodynamic Stability of Folded Proteins Against Mutations. Physical Review Letters 79(18), pp.3530-3533.

Caetano-Anolles, G. Wang, M. Caetano-Anolles, D. and Mittenthal, JE. (2009) The Origin, Evolution and Structure of the Protein World. The Biochemical Journal 417(3), pp. 621-637.

Campanacci, V. Nurizzo, D. Spinelli, S. Valencia, C. Tegoni, M. Cambillau, C. (2004). The crystal structure of the Escherichia coli lipocalin Blc suggests a possible role in phospholipid binding. FEBS Lett. 562(1-3), pp. 183-188.

Chothia, C. and Gough, J. (2009) Genomic and Structural Aspects of Protein Evolution. The Biochemical Journal 419(1), pp. 15-28.

Chu, CK. Feng LL. and Wouters. MA. (2005) Comparison of sequence and structure-based datasets for nonredundant structural data mining. Proteins 60(4), pp. 577-583.

Ciccarelli et al. (2006) Toward Automatic Reconstruction of a Highly Resolved Tree of Life. Science, Volume 311, Pages 1283-1287.

Csaba, G. Birzele, F. and Zimmer, R. (2009) Systematic comparison of SCOP and CATH: a new gold standard for protein structure analysis. BMC Structural Biology, 9 (23).

Cuff et al.(2009) The CATH Classification revisited–architectures reviewed and new ways to characterize structural divergence in superfamilies. Nucleic Acids Research, Volume 38, Pages D310–D314.

Forslund, K. Henricson, A. Hollich, V. and Sonnhammer, EL (2008) Domain Tree-Based Analysis of Protein Architecture Evolution. Molecular Biology and Evolution 25(2) , pp.254-264.

Forwood, JK. (2010) Quantitative structural analysis of importin-β flexibility: Paradigm for solenoid protein structures. Structure 18(9), pp.1171-1183.

Gerstein, M and Hegyi, H. (1998) Comparing genomes in terms of protein structure: Surveys of a finite parts list. FEMS Microbioliology Review 22 (4), pp. 277–304.

Giangreco, I. et al. (2011) Insights into the Complex Formed by Matrix Metalloproteinase-2 and Alloxan Inhibitors: Molecular Dynamics Simulations and Free Energy Calculations. PLoS ONE 6 (10), e25597.

Gough, J. Chothia, C. (2002) SUPERFAMILY: HMMs representing all proteins of known structure. SCOP sequence searches, alignments and genome assignments. Nucleic Acids Research 30 (1), pp. 268-272.

Greene, LH. Lewis, TE. Addou, S. Cuff, A. Dallman, T. Dibley, M. Redfern, O. Pearl, F. Nambudiry, R. Reid, A. Sillitoe, I. Yeats, C. Thornton, JM. Orengo, CA. (2007). The CATH domain structure database: new protocols and classification levels give a more comprehensive resource for exploring evolution. Nucleic Acids Research 35, D291-D297.

Gribaldo, S. and Brochier-Armanet, C. (2006) The origin and evolution of Archaea: a state of the art Phil Trans R Soc B 361: 1007-1022.

Groves, MR. Hanlon, N. Turowski, P. Hemmings, BA. and Barford, D. (1999). The structure of the protein phosphatase 2A PR65/A subunit reveals the conformation of its 15 tandemly repeated HEAT motifs. Cell 96(1) , pp.99-110.

Hartman et al. (2005) The archaeal origins of the eukaryotic translational system. Archaea, Volume 2, Pages 1-9.

Harrison et al. (2002) Quantifying the similarities within fold space. Journal of Molecular Biology, Volume 323, Pages 909–926.

Heinemann, U. Illing, G. Oschkinat, H. (2001) High-throughput three-dimensional protein structure determination. Current Opinion in Biotechnology, 12(4), pp.348-354.

Holm, L. and Sander, C. (1994) Parser for protein folding units. Proteins, 19(3), pp. 256-268.

Holm, L. Ouzounis, C. Sander, C. Tuparev, G. Vriend, G. (1992) A database of protein structure families with common folding motifs. Protein Science 1(12), pp. 1691-1698.

Holm, L. Sander, C. (1998) Touring protein fold space with dali/fssp. Nucleic Acids Research 26(1), pp. 316-319.

Kendrew, JC. Bodo, G. Dintzis, HM. Parrish, RG. Wyckoff, H. Philips, DC. (1958) A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. Nature 181(4610), pp. 662-666.

Kim KM, Caetano-Anollés G. (2012) The evolutionary history of protein fold families and proteomes confirms that the archaeal ancestor is more ancient than the ancestors of other superkingdoms. BMC Evolutionary Biology.

Kim, KM. and Caetano-Anollés G (2011) The proteomic complexity and rise of the primordial ancestor of diversified life. BMC Evolutionary Biology Vol.11 no.140 doi: 10.1186/1471-2148-11-140.

Kobe, B. and Kajava, AV. (2000) When protein folding is simplified to protein coiling: the continuum of solenoid protein structures. Trends in Biochemical Science 25 (10), pp. 509-515.

Kobe, B. Gleichmann, T. Horne, J. Jennings, IG. Scotney, PD. and Teh, T. (1999). Turn up the HEAT. Structure 7 (5), R91-R97.

Kobe, B and Deisenhofer, J. (1993) Crystal structure of porcine ribonuclease inhibitor, a protein with leucine-rich repeats. Nature, 366 (6457), pp.751-756.

Lecompte, O. Ripp, R. Thierry, JC. Moras, D. and Poch, O. (2002) Comparative analysis of ribosomal proteins in complete genomes: an example of reductive evolution at the domain scale. Nucleic Acids Research 30(24), pp.5382-5390 .

Lee, D., Grant, A., Marsden, R. L., Orengo, C., May 2005. Identification and distribution of protein families in 120 completed genomes using gene3d. Proteins 59 (3), pp. 603-615.

Lee, JC. Herman, P. (2011) Structural and functional energetic linkages in allosteric regulation of muscle pyruvate kinase. Methods in Enzymology 488, pg. 185-217.

Lees, J. Yeats, C. Redfern, O. Clegg, A. and Orengo, C. (2010) Gene3D: Merging Structure and Function for a Thousand Genomes. Nucleic Acids Research 38(1), D296-300.

Letunic, I. and Bork, P. (2011) Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy. Nucleic Acids Research 39, W475-W478.

Lienau et al. (2011) The mega-matrix tree of life: using genome-scale horizontal gene transfer and sequence evolution data as information about the vertical history of life, Journal of Cladistics, Volume 27, Issue 4, pages 417–427.

Liu, J. Montelione, GT. Rost, B. (2007) Novel leverage of structural genomics. Nature Biotechnology, 25(8), pp. 849-851.

Madera, M. Gough, J. (2002). A comparison of profile hidden markov model procedures for remote homology detection. Nucleic Acids Research 30(19), pp. 4321-4328.

Michie, AD. Orengo, CA. Thornton, JM. (1996) Analysis of domain structural class using an automated class assignment protocol. Journal of Molecular Biology 262(2), pp.168-185.

Mike Steel, Andy McKenzie, Properties of phylogenetic trees generated by Yule-type speciation models, Mathematical Biosciences, Volume 170, Issue 1, March 2001, Pages 91-112, ISSN 0025-5564, 10.1016/S0025-5564(00)00061-4.

Murzin, A. Brenner, S. Hubbard, T. Chothia, C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. Journal of Molecular Biology 247(4), pp. 536-540.

Murzin, AG. Brenner, SE. Hubbard, T and Clothia, C. (1995) SCOP: A structural classification of proteins for the investigation of sequences and structures. Journal of Molecular Biololgy 247 (4), pp.536–540.

Ogrunc, M. Becker, DF. Ragsdale, SW. and Sancar, A. (1998) Nucleotide Excision Repair in the Third Kingdom. Journal of Bacteriology 180(21) , pp. 5796-5798.

Orengo, CA. et al. (1997) CATH — a hierarchic classification of protein domain structures. Structure, 5 (8), pp.1093-1108.

Peifer, M. Berg, S. and Reynolds, AB. (1994) A repeating amino acid motif shared by proteins with diverse cellular roles. Cell 76(5), pp.789–791.

Rao, ST. Rossmann, MG. (1973) Comparison of super-secondary structures in proteins. Journal of Molecular Biology 76 (2), pp.241-56.

Rees, DC. Komiya, H. Yeates, TO. Allen, JP. Feher, G. (1989) The bacterial photosynthetic reaction center as a model for membrane proteins. Annual review of biochemistry 58, pp. 607-633.

Reeves et al. (2006) Structural diversity of domain superfamilies in the CATH database. Journal of Molecular Biology, Volume 360, Pages 725-741.

Richardson, JS. (1981) The anatomy and taxonomy of protein structure. Adv Protein Chem 34, 167–339.

Rossmann, MG. Argos, P. (1976) Exploring structural homology of proteins. Journal of Molecular Biology 105(1), pp. 75-95.

Scheraga, HA. Wedemeyer, WJ. Welker, E. (2001) Bovine pancreatic ribonuclease A: oxidative and conformational folding studies. Methods Enzymology 341, pp.189-221.

Service, RF. (2008) Protein structure initiative: phase 3 or phase out. Science, 319(5870), pp. 1610-1613.

Siddiqui, AS. and Barton, GJ. (1995) Continuous and discontinuous domains: an algorithm for the automatic generation of reliable protein domain definitions. Protein Science, 4(5), pp. 872-884.

Stargell, LA. et al. (2001) Transcriptional activity of the TFIIA four-helix bundle in vivo. Proteins 43(2), pp. 227-232.

Sun F-J, Caetano-Anollés G. (2008) The origin and evolution of tRNA inferred from phylogenetic analysis of structure. The Journal of Molecular Evolution, Volume 66, Pages 21–35.

Sun F-J, Caetano-Anollés G. (2009) The evolutionary history of the structure of 5S ribosomal RNA. The Journal of Molecular Evolution, Volume 69, Pages 430–443.

Sun F-J, Caetano-Anollés G. (2010) The ancient history of the structure of ribonuclease P and the early origins of Archaea. BMC Bioinformatics.

Sun F-J et al. (2006) Common evolutionary trends for tRNA-derived SINE RNA structures. Trends in Genetics Volume 23, Pages 26–33.

Swindells, MB. MacArthur, MW. Thornton, JM. (1995) Intrinsic phi, psi propensities of amino acids, derived from the coil regions of known structures. Nature Structure Biology 2(7), pp. 596-603.

Swofford DL: Phylogenetic analysis using parsimony and other program (PAUP*), ver. 4.0b10. Sinauer, Sunderland, MA; 2002.

Taylor, RW. Aszodi, A. (2005) Protein Geometry, Classification, Topology and Symmetry Institute of Physics Publishing Bristol and Philadelphia.

Wallin, E, Tsukihara, T. Yoshikawa, S. Von Heijne, G. Elofsson, A. (1997) Architecture of helix bundle membrane proteins: an analysis of cytochrome c oxidase from bovine mitochondria. Protein Science 6(4), pp. 808-815.

Wang, M. et al. (2011) A universal molecular clock of protein folds and its power in tracing the early history of aerobic metabolism and planet oxygenation. Molecular Biology and Evolution 28 (1), pp. 567-582.

Wang, M. Yafremava, LS. Caetano-Anollés, D. Mittenthal JE. and Caetano-Anollés, G. (2007) Reductive evolution of architectural repertoires in proteomes and the birth of the tripartite world Genome Research, 17 (11), pp. 1572-1585.

Weiss, MS. Abele, U. Weckesser, J. Welte, W. Schiltz, E. Schulz, GE. (1991) Molecular architecture and electrostatic properties of a bacterial porin. Science 254 (5038), pp. 1627-1630.

Wimley, WC. (2003) The versatile beta-barrel membrane protein. Current opinion in structural biology Vol.13 no.4, pp. 404-411.

Woese, CR. and Fox, GE. (1977) Phylogenetic Structure of the Prokaryotic Domain: The Primary Kingdoms. Proceedings of the National Academy of Sciences of the United States of America 74(11), pp.5088-5090.

Worth, CL. Gong, S.and Blundell, TL. (2009) Structural and functional contraints in the evolution of protein families. Nature Review Molecular Cell Biology, 10(10), pp. 709-720.

Xue et al. (2003) Transfer RNA paralogs: evidence for genetic code-amino acid biosynthesis coevolution and an archaeal root of life. Gene, Volume 310, Pages 59–66.

Yang, S. Bourne PE. (2009) The Evolutionary History of Protein Domains Viewed by Species Phylogeny. PLoS One 4(12) , e8378.

Yeats, C. Lees, J. Carter, P. Sillitoe, I. and Orengo, C. (2011) The Gene3D Web Services: a platform for identifying, annotating and comparing structural domains in protein sequences Nucleic Acids Research, 39 (7), W546-W550.

APPENDIX A

SUPPLEMENTARY DATA

Table A.1 List of 492 organism with their genome id and genome names. A, B and E letters at the end of genome name refers to superkingdoms the Archaea, the Bacteria and the Eukarya respectively.

Genome ID	***Genome Name***
00	Campylobacter hominis_B
01	Polaromonas naphthalenivorans_B
02	Metallosphaera sedula_A
03	Clostridium beijerinckii_B
04	Borrelia afzelii_B
09	Pelotomaculum thermopropionicum_B
0C	Anaplasma marginale_B
0G	Sulfolobus islandicus_A
0K	Escherichia coli_B
0L	Streptococcus suis_B
0N	Laribacter hongkongensis_B
0P	Gluconacetobacter diazotrophicus_B
0R	Atopobium parvulum_B
0S	Streptococcus equi_B
0V	Desulfovibrio desulfuricans_B
0X	Thermococcus sibiricus_A
0Z	Clostridium cellulolyticum_B
11	Methanosaeta thermophila_A
12	Clostridium novyi_B
13	Lactobacillus gasseri_B
14	Methylibium petroleiphilum_B
16	Clostridium kluyveri_B
17	Rhizobium leguminosarum_B
18	Saccharopolyspora erythraea_B
1B	Deinococcus deserti_B
1E	Brucella melitensis_B
1H	Anaerococcus prevotii_B
1N	Clostridium botulinum_B
1O	Eggerthella lenta_B
1P	Bacillus cereus_B
1Q	Saccharomonospora viridis_B
1R	Mycobacterium tuberculosis_B
1S	Acetobacter pasteurianus_B
1U	Streptococcus pneumoniae_B
1V	Macrococcus caseolyticus_B
1W	Methylotenera mobilis_B
1X	Helicobacter pylori_B

Table A.1 (contd.)

Genome ID	*Genome Name*
1Z	Actinosynnema mirum_B
23	Methanoculleus marisnigri_A
27	Paracoccus denitrificans_B
28	Mycobacterium vanbaalenii_B
2A	Klebsiella pneumoniae_B
2B	Variovorax paradoxus_B
2D	Dyadobacter fermentans_B
2E	Kosmotoga olearia_B
2G	Kytococcus sedentarius_B
2K	Desulfohalobium retbaense_B
2L	Acidobacterium capsulatum_B
2O	Teredinibacter turnerae_B
2P	Desulfotomaculum acetoxidans_B
2Q	Micrococcus luteus_B
2V	Nakamurella multipartita_B
2X	Jonesia denitrificans_B
2Y	Burkholderia glumae_B
2Z	Ralstonia pickettii_B
30	Burkholderia vietnamiensis_B
32	Clavibacter michiganensis_B
33	Aeromonas hydrophila_B
34	Pyrobaculum islandicum_A
35	Campylobacter fetus_B
37	Cytophaga hutchinsonii_B
3B	Caulobacter crescentus_B
3C	Burkholderia pseudomallei_B
3E	Rhodococcus opacus_B
3G	Brachybacterium faecium_B
3I	Dickeya zeae_B
3J	Brachyspira hyodysenteriae_B
3L	Aggregatibacter aphrophilus_B
3M	Tolumonas auensis_B
3N	Eubacterium eligens_B
3P	Methanocaldococcus fervens_A
3Q	Rhodobacter sphaeroides_B
3R	Halorhabdus utahensis_A
3W	Desulfobacterium autotrophicum_B
3Y	Pedobacter heparinus_B
40	Ochrobactrum anthropi_B
44	Aeromonas salmonicida_B
45	Mycoplasma agalactiae_B

Table A.1 (contd.)

Genome ID	*Genome Name*
49	Lactobacillus reuteri_B
4A	Alicyclobacillus acidocaldarius_B
4B	Halorubrum lacusprofundi_A
4C	Slackia heliotrinireducens_B
4D	Catenulispora acidiphila_B
4H	Kangiella koreensis_B
4I	Campylobacter lari_B
4L	Lactobacillus plantarum_B
4M	Thermomicrobium roseum_B
4T	Pseudomonas fluorescens_B
4U	Listeria monocytogenes_B
4V	Flavobacteriaceae bacterium_B
4X	Halomicrobium mukohataei_A
4Y	Mycobacterium leprae_B
52	Lactobacillus brevis_B
54	Hyphomonas neptunium_B
56	Arthrobacter aurescens_B
57	Methanococcus vannielii_A
58	Kineococcus radiotolerans_B
5B	Salmonella enterica_B
5D	Halothermothrix orenii_B
5E	Anaerocellum thermophilum_B
5O	Vibrio cholerae_B
5R	Lactobacillus rhamnosus_B
5S	Staphylococcus carnosus_B
5U	Beutenbergia cavernae_B
5V	Thermococcus gammatolerans_A
5Y	Chitinophaga pinensis_B
60	Sorangium cellulosum_B
61	Salinispora arenicola_B
62	Bacillus weihenstephanensis_B
63	Ignicoccus hospitalis_A
65	Vibrio harveyi_B
68	Alkaliphilus oremlandii_B
69	Cronobacter sakazakii_B
6A	Methylobacterium extorquens_B
6B	Gemmatimonas aurantiaca_B
6C	Nautilia profundicola_B
6D	Desulfomicrobium baculatum_B
6E	Arthrobacter chlorophenolicus_B
6H	Mycoplasma hominis_B

Table A.1 (contd.)

Genome ID	*Genome Name*
6Q	Rothia mucilaginosa_B
6R	Xanthomonas albilineans_B
6T	Staphylococcus lugdunensis_B
6U	Lactococcus lactis_B
6W	Listeria seeligeri_B
74	Rickettsia akari_B
75	Azorhizobium caulinodans_B
76	Burkholderia multivorans_B
77	Roseiflexus castenholzii_B
79	Lactobacillus helveticus_B
7D	Streptococcus gallolyticus_B
7E	Streptomyces scabiei_B
7F	Geodermatophilus obscurus_B
7I	Xylanimonas cellulosilytica_B
7N	Deferribacter desulfuricans_B
7O	Thermocrinis albus_B
7R	Erwinia pyrifoliae_B
7S	Aggregatibacter actinomycetemcomitans_B
7V	Clostridium difficile_B
7W	Archaeoglobus profundus_A
7Z	Sphaerobacter thermophilus_B
81	Acaryochloris marina_B
82	Thermoanaerobacter pseudethanolicus_B
83	Leptospira biflexa_B
84	Microcystis aeruginosa_B
88	Shewanella woodyi_B
8C	Lactobacillus johnsonii_B
8G	Gordonia bronchialis_B
8H	Erwinia amylovora_B
8I	Zymomonas mobilis_B
8J	Salinibacter ruber_B
8L	Veillonella parvula_B
8M	Edwardsiella tarda_B
8P	Sebaldella termitidis_B
8Q	Legionella longbeachae_B
8T	Fibrobacter succinogenes_B
8U	Streptosporangium roseum_B
90	Burkholderia phymatum_B
94	Elusimicrobium minutum_B
98	Kocuria rhizophila_B
9A	Staphylococcus aureus_B

Table A.1 (contd.)

Genome ID	*Genome Name*
9C	Thermanaerovibrio acidaminovorans_B
9H	Stackebrandtia nassauensis_B
9L	Hydrogenobacter thermophilus_B
9O	Aciduliprofundum boonei_A
9Q	Methanobrevibacter ruminantium_A
9W	alpha proteobacterium_B
9Y	Pantoea ananatis_B
9Z	Clostridiales genomosp._B
C1	Callithrix jacchus_E
CD	Candida dubliniensis_E
PC	Penicillium chrysogenum_E
SS	Sus scrofa_E
TV	Trichophyton verrucosum_E
a5	Aspergillus niger_E
a7	Aspergillus clavatus_E
a8	Aspergillus oryzae_E
aA	Pirellula staleyi_B
aE	Haliangium ochraceum_B
aG	Haloferax volcanii_A
aL	Meiothermus ruber_B
aN	Bacillus pseudofirmus_B
aP	Streptococcus mitis_B
aQ	Ferroglobus placidus_A
aT	Chlamydia trachomatis_B
au	Agrobacterium tumefaciens_B
av	Mycobacterium avium_B
ax	Aedes aegypti_E
az	Aromatoleum aromaticum_B
b1	Baumannia cicadellinicola_B
b2	Bacillus anthracis_B
b3	Brucella abortus_B
b4	Burkholderia xenovorans_B
b5	Burkholderia thailandensis_B
b6	Burkholderia cenocepacia_B
bb	Borrelia burgdorferi_B
be	Bordetella pertussis_B
bh	Bacillus halodurans_B
bi	Burkholderia mallei_B
bj	Bradyrhizobium japonicum_B
bl	Bifidobacterium longum_B
bn	Buchnera aphidicola_B

Table A.1 (contd.)

Genome ID	*Genome Name*
bo	Bordetella bronchiseptica_B
bp	Bordetella parapertussis_B
bq	Bartonella quintana_B
br	Brucella suis_B
bs	Bacillus subtilis_B
bt	Bacteroides thetaiotaomicron_B
bv	Bos taurus_E
c0	Ciona savignyi_E
c2	Chlamydophila pneumoniae_B
c3	Corynebacterium glutamicum_B
c4	Chlamydophila abortus_B
c5	Chlorobium chlorochromatii_B
c6	Chlamydophila felis_B
c8	Chromohalobacter salexigens_B
c9	Carboxydothermus hydrogenoformans_B
ca	Clostridium acetobutylicum_B
cf	Cryptococcus neoformans_E
ch	Chlorobium tepidum_B
cl	Caenorhabditis elegans_E
cm	Chlamydia muridarum_B
co	Corynebacterium efficiens_B
cv	Cryptosporidium parvum_E
cw	Caenorhabditis briggsae_E
d0	Saccharophagus degradans_B
d1	Shigella dysenteriae_B
d2	Desulfitobacterium hafniense_B
d4	Deinococcus geothermalis_B
d5	Dasypus novemcinctus_E
da	Danio rerio_E
dd	Drosophila melanogaster_E
dj	Dechloromonas aromatica_B
do	Drosophila pseudoobscura_E
dp	Desulfotalea psychrophila_B
dr	Deinococcus radiodurans_B
dt	Dictyostelium discoideum_E
dv	Desulfovibrio vulgaris_B
e9	Stenotrophomonas maltophilia_B
ee	Echinops telfairi_E
ef	Enterococcus faecalis_B
eg	Ehrlichia chaffeensis_B
eh	Ehrlichia ruminantium_B

Table A.1 (contd.)

Genome ID	*Genome Name*
ej	Erwinia tasmaniensis_B
ek	Erinaceus europaeus_E
el	Ehrlichia canis_B
em	Leishmania major_E
ep	Staphylococcus epidermidis_B
eq	Equus caballus_E
er	Pectobacterium atrosepticum_B
et	Dehalococcoides ethenogenes_B
eu	Encephalitozoon cuniculi_E
ev	Chlorobaculum parvum_B
ey	Erythrobacter litoralis_B
f7	Borrelia duttonii_B
fb	Plasmodium berghei_E
fd	Desulfurococcus kamchatkensis_A
fe	Felis catus_E
fj	Methylobacterium populi_B
fl	Bacteroides fragilis_B
fn	Fusobacterium nucleatum_B
fp	Chlorobium limicola_B
fs	Shigella flexneri_B
ft	Francisella tularensis_B
fu	Methanosarcina barkeri_A
fw	Plasmodium knowlesi_E
fy	Plasmodium chabaudi_E
g5	Bifidobacterium animalis_B
g7	Borrelia recurrentis_B
ga	Borrelia garinii_B
gb	Spermophilus tridecemlineatus_E
gc	Gasterosteus aculeatus_E
gf	Giardia lamblia_E
gg	Gallus gallus_E
gi	Aspergillus terreus_E
gk	Geobacillus kaustophilus_B
gl	Candida glabrata_E
gm	Geobacter metallireducens_B
go	Ashbya gossypii_E
gq	Aspergillus flavus_E
gs	Geobacter sulfurreducens_B
gt	Guillardia theta_E
gu	Cavia porcellus_E
gv	Gloeobacter violaceus_B

Table A.1 (contd.)

Genome ID	*Genome Name*
gx	Gorilla gorilla_E
h3	Coprothermobacter proteolyticus_B
h5	Shewanella piezotolerans_B
h6	Thermococcus onnurineus_A
ha	Pseudoalteromonas haloplanktis_B
hc	Hahella chejuensis_B
hd	Haemophilus ducreyi_B
he	Photobacterium profundum_B
hg	Chaetomium globosum_E
hh	Helicobacter hepaticus_B
hi	Haemophilus influenzae_B
hl	Helicobacter acinonychis_B
hm	Haloarcula marismortui_A
ho	Bartonella henselae_B
hs	Homo sapiens_E
hw	Shewanella denitrificans_B
ib	Leishmania braziliensis_E
ih	Tarsius syrichta_E
il	Idiomarina loihiensis_B
io	Microcebus murinus_E
ir	Paramecium tetraurelia_E
is	Ciona intestinalis_E
ix	Dictyoglomus turgidum_B
j0	Natranaerobius thermophilus_B
j5	Methylobacterium chloromethanicum_B
j6	Anoxybacillus flavithermus_B
jb	Chloroflexus aggregans_B
jj	Oligotropha carboxidovorans_B
jk	Corynebacterium jeikeium_B
jn	Proteus mirabilis_B
jt	Thermosipho africanus_B
jw	Geobacter bemidjiensis_B
jx	Methanosphaerula palustris_A
k1	Aliivibrio salmonicida_B
k3	Dictyoglomus thermophilum_B
k7	Thermodesulfovibrio yellowstonii_B
kb	Alteromonas macleodii_B
kd	Escherichia fergusonii_B
ke	Methylocella silvestris_B
kf	Ureaplasma urealyticum_B
kg	Rhodospirillum centenum_B

Table A.1 (contd.)

Genome ID	*Genome Name*
kj	Acidithiobacillus ferrooxidans_B
kk	Phenylobacterium zucineum_B
kl	Kluyveromyces lactis_E
km	Haemophilus parasuis_B
kn	Vibrio splendidus_B
ko	Bacillus thuringiensis_B
ku	Burkholderia phytofirmans_B
ld	Lactobacillus delbrueckii_B
lf	Bacillus licheniformis_B
lh	Leishmania infantum_E
li	Listeria innocua_B
lk	Loxodonta africana_E
ln	Lawsonia intracellularis_B
lr	Leptospira interrogans_B
ls	Lactobacillus sakei_B
lt	Lactobacillus acidophilus_B
lu	Myotis lucifugus_E
lv	Lactobacillus salivarius_B
lw	Colwellia psychrerythraea_B
lx	Leifsonia xyli_B
ly	Lodderomyces elongisporus_E
m0	Mycoplasma mobile_B
m2	Methanococcus maripaludis_A
m3	Mannheimia succiniciproducens_B
m4	Methanosphaera stadtmanae_A
m5	Magnetospirillum magneticum_B
m6	Methanospirillum hungatei_A
m7	Methylobacillus flagellatus_B
m8	Moorella thermoacetica_B
m9	Methanococcoides burtonii_A
ma	Methanosarcina acetivorans_A
mc	Mycobacterium bovis_B
md	Methanothermobacter thermautotrophicus_A
me	Mycoplasma penetrans_B
mf	Mesoplasma florum_B
mg	Mycoplasma genitalium_B
mj	Methanocaldococcus jannaschii_A
mk	Mesorhizobium loti_B
mm	Mus musculus_E
mn	Methanopyrus kandleri_A
mp	Mycoplasma pneumoniae_B

Table A.1 (contd.)

Genome ID	*Genome Name*
mq	Mycoplasma pulmonis_B
mt	Methylococcus capsulatus_B
my	Mycoplasma gallisepticum_B
mz	Methanosarcina mazei_A
na	Nanoarchaeum equitans_A
nb	Nitrobacter hamburgensis_B
ne	Nitrosomonas europaea_B
nf	Nocardia farcinica_B
nh	Neosartorya fischeri_E
ni	Neisseria gonorrhoeae_B
nl	Nitrosospira multiformis_B
nn	Neisseria meningitidis_B
np	Natronomonas pharaonis_A
nr	Nitrosococcus oceani_B
ns	Neurospora crassa_E
nt	Thiobacillus denitrificans_B
nu	Theileria annulata_E
nv	Novosphingobium aromaticivorans_B
nw	Nematostella vectensis_E
o0	Babesia bovis_E
ob	Otolemur garnettii_E
of	Pongo pygmaeus_E
oh	Ornithorhynchus anatinus_E
oi	Oceanobacillus iheyensis_B
ok	Oryctolagus cuniculus_E
ol	Oryzias latipes_E
on	Shigella sonnei_B
op	Monodelphis domestica_E
oq	Ochotona princeps_E
os	Oryza sativa_E
ou	Ostreococcus tauri_E
ov	Monosiga brevicollis_E
ox	Gluconobacter oxydans_B
oz	Ostreococcus lucimarinus_E
p1	Prochlorococcus marinus_B
p3	Picrophilus torridus_A
p8	Pelobacter carbinolicus_B
pa	Pseudomonas aeruginosa_B
pb	Pyrococcus abyssi_A
pd	Photorhabdus luminescens_B
pg	Porphyromonas gingivalis_B

Table A.1 (contd.)

Genome ID	*Genome Name*
ph	Pyrococcus horikoshii_A
pj	Pseudomonas syringae_B
pl	Plasmodium falciparum_E
pq	Pseudomonas entomophila_B
ps	Pseudomonas putida_B
pu	Pyrococcus furiosus_A
pv	Theileria parva_E
px	Pseudoalteromonas atlantica_B
py	Plasmodium yoelii_E
pz	Psychrobacter cryohalolentis_B
r3	Rickettsia bellii_B
rb	Rhodospirillum rubrum_B
rc	Rickettsia conorii_B
rd	Rhodopseudomonas palustris_B
rf	Rickettsia felis_B
ri	Psychrobacter arcticus_B
rl	Ralstonia eutropha_B
rm	Cryptosporidium hominis_E
rn	Rattus norvegicus_E
rp	Rickettsia prowazekii_B
rs	Ralstonia solanacearum_B
rt	Rickettsia typhi_B
ru	Macaca mulatta_E
rw	Rubrobacter xylanophilus_B
rx	Rhodoferax ferrireducens_B
rz	Rhizobium etli_B
s5	Streptococcus agalactiae_B
s6	Streptococcus mutans_B
s9	Streptomyces avermitilis_B
sb	Symbiobacterium thermophilum_B
sc	Saccharomyces cerevisiae_E
sd	Streptococcus thermophilus_B
sf	Streptomyces coelicolor_B
sg	Ruegeria pomeroyi_B
sm	Sinorhizobium meliloti_B
sq	Shigella boydii_B
sv	Sulfolobus tokodaii_A
t0	Thermus thermophilus_B
ta	Thermoplasma acidophilum_A
tc	Thiomicrospira crunogena_B
td	Treponema denticola_B

Table A.1 (contd.)

Genome ID	*Genome Name*
te	Clostridium tetani_B
tf	Thermobifida fusca_B
ti	Sulfurimonas denitrificans_B
tk	Thermococcus kodakarensis_A
tn	Tetraodon nigroviridis_E
to	Takifugu rubripes_E
tw	Tropheryma whipplei_B
tz	Tupaia belangeri_E
ue	Toxoplasma gondii_E
um	Ustilago maydis_E
ut	Tursiops truncatus_E
uu	Ureaplasma parvum_B
uz	Trypanosoma cruzi_E
va	Anabaena variabilis_B
vb	Vibrio vulnificus_B
vf	Vibrio fischeri_B
vi	Chromobacterium violaceum_B
vn	Procavia capensis_E
vp	Vibrio parahaemolyticus_B
vr	Pteropus vampyrus_E
vw	Vanderwaltozyma polyspora_E
vx	Plasmodium vivax_E
wb	Wigglesworthia glossinidia_B
wi	Nitrobacter winogradskyi_B
ws	Wolinella succinogenes_B
x2	Streptococcus pyogenes_B
x3	Synechococcus elongatus_B
x4	Bacillus clausii_B
x7	Legionella pneumophila_B
x9	Mycoplasma hyopneumoniae_B
xc	Xanthomonas axonopodis_B
xd	Xanthomonas campestris_B
xf	Xylella fastidiosa_B
xo	Xanthomonas oryzae_B
xp	Pan troglodytes_E
xr	Sorex araneus_E
yc	Mycoplasma synoviae_B
yi	Mycoplasma capricolum_B
yl	Yarrowia lipolytica_E
yp	Yersinia pestis_B
yr	Yersinia pseudotuberculosis_B

Table A.1 (contd.)

Genome ID	*Genome Name*
yt	Syntrophus aciditrophicus_B
za	Sulfolobus acidocaldarius_A
zh	Staphylococcus haemolyticus_B
zt	Staphylococcus saprophyticus_B

Table A.2 List of 295 FL proteomes used to construct Proteome trees.

Genome ID	*Phylum*	*Superkingdom*	*NCBI Taxonomy ID*
cf	Fungi	Eukaryota	214684
hg	Fungi	Eukaryota	306901
ns	Fungi	Eukaryota	367110
nh	Fungi	Eukaryota	331117
gi	Fungi	Eukaryota	341663
a8	Fungi	Eukaryota	5062
a5	Fungi	Eukaryota	425011
gq	Fungi	Eukaryota	332952
a7	Fungi	Eukaryota	344612
ly	Fungi	Eukaryota	379508
yl	Fungi	Eukaryota	284591
vw	Fungi	Eukaryota	436907
gl	Fungi	Eukaryota	284593
go	Fungi	Eukaryota	284811
sc	Fungi	Eukaryota	4932
kl	Fungi	Eukaryota	284590
hs	Metazoa	Eukaryota	9606
xp	Metazoa	Eukaryota	9598
gx	Metazoa	Eukaryota	9593
of	Metazoa	Eukaryota	9600
ru	Metazoa	Eukaryota	9544
C1	Metazoa	Eukaryota	9483
ob	Metazoa	Eukaryota	30611
io	Metazoa	Eukaryota	30608
ih	Metazoa	Eukaryota	9478
rn	Metazoa	Eukaryota	10116
mm	Metazoa	Eukaryota	10090
gb	Metazoa	Eukaryota	43179
gu	Metazoa	Eukaryota	10141
ok	Metazoa	Eukaryota	9986
oq	Metazoa	Eukaryota	9978
tz	Metazoa	Eukaryota	37347
SS	Metazoa	Eukaryota	9823
bv	Metazoa	Eukaryota	9913
ut	Metazoa	Eukaryota	9739
fe	Metazoa	Eukaryota	9685
eq	Metazoa	Eukaryota	9796
lu	Metazoa	Eukaryota	59463
vr	Metazoa	Eukaryota	132908

Table A.2 (contd.)

Genome ID	*Phylum*	*Superkingdom*	*NCBI Taxonomy ID*
xr	Metazoa	Eukaryota	42254
ek	Metazoa	Eukaryota	9365
vn	Metazoa	Eukaryota	9813
lk	Metazoa	Eukaryota	9785
ee	Metazoa	Eukaryota	9371
d5	Metazoa	Eukaryota	9361
op	Metazoa	Eukaryota	13616
oh	Metazoa	Eukaryota	9258
gg	Metazoa	Eukaryota	9031
da	Metazoa	Eukaryota	7955
gc	Metazoa	Eukaryota	69293
ol	Metazoa	Eukaryota	8090
tn	Metazoa	Eukaryota	99883
to	Metazoa	Eukaryota	31033
c0	Metazoa	Eukaryota	51511
is	Metazoa	Eukaryota	7719
do	Metazoa	Eukaryota	46245
dd	Metazoa	Eukaryota	7227
ax	Metazoa	Eukaryota	7159
cl	Metazoa	Eukaryota	6239
cw	Metazoa	Eukaryota	6238
nw	Metazoa	Eukaryota	45351
ir	Protista	Eukaryota	5888
os	Plantae	Eukaryota	39947
oz	Plantae	Eukaryota	436017
ou	Plantae	Eukaryota	70448
2L	Acidobacteria	Bacteria	240015
g5	Actinobacteria	Bacteria	580050
58	Actinobacteria	Bacteria	266940
4D	Actinobacteria	Bacteria	479433
9H	Actinobacteria	Bacteria	446470
2V	Actinobacteria	Bacteria	479431
7F	Actinobacteria	Bacteria	526225
tf	Actinobacteria	Bacteria	269800
8U	Actinobacteria	Bacteria	479432
s9	Actinobacteria	Bacteria	227882
sf	Actinobacteria	Bacteria	100226
18	Actinobacteria	Bacteria	405948
61	Actinobacteria	Bacteria	391037
3E	Actinobacteria	Bacteria	632772
nf	Actinobacteria	Bacteria	247156

Table A.2 (contd.)

Genome ID	*Phylum*	*Superkingdom*	*NCBI Taxonomy ID*
av	Actinobacteria	Bacteria	262316
28	Actinobacteria	Bacteria	350058
co	Actinobacteria	Bacteria	196164
c3	Actinobacteria	Bacteria	196627
5U	Actinobacteria	Bacteria	471853
2X	Actinobacteria	Bacteria	471856
3G	Actinobacteria	Bacteria	446465
6E	Actinobacteria	Bacteria	452863
56	Actinobacteria	Bacteria	290340
2Q	Actinobacteria	Bacteria	596312
4C	Actinobacteria	Bacteria	471855
0R	Actinobacteria	Bacteria	521095
rw	Actinobacteria	Bacteria	266117
7O	Aquificae	Bacteria	638303
9L	Aquificae	Bacteria	608538
37	Bacteroidetes	Bacteria	269798
8J	Bacteroidetes	Bacteria	309807
5Y	Bacteroidetes	Bacteria	485918
3Y	Bacteroidetes	Bacteria	485917
4V	Bacteroidetes	Bacteria	531844
fp	Chlorobi	Bacteria	290315
ev	Chlorobi	Bacteria	517417
ch	Chlorobi	Bacteria	194439
et	Chloroflexi	Bacteria	243164
4M	Chloroflexi	Bacteria	309801
7Z	Chloroflexi	Bacteria	479434
77	Chloroflexi	Bacteria	383372
jb	Chloroflexi	Bacteria	326427
gv	Cyanobacteria	Bacteria	251221
81	Cyanobacteria	Bacteria	329726
p1	Cyanobacteria	Bacteria	59922
va	Cyanobacteria	Bacteria	240292
84	Cyanobacteria	Bacteria	449447
7N	Deferribacteres	Bacteria	639282
1B	Deinococcus-Thermus	Bacteria	546414
d4	Deinococcus-Thermus	Bacteria	319795
dr	Deinococcus-Thermus	Bacteria	243230
aL	Deinococcus-Thermus	Bacteria	504728
t0	Deinococcus-Thermus	Bacteria	262724
ix	Dictyoglomi	Bacteria	515635
k3	Dictyoglomi	Bacteria	309799

Table A.2 (contd.)

Genome ID	*Phylum*	*Superkingdom*	*NCBI Taxonomy ID*
8T	Fibrobacteres	Bacteria	59374
j0	Firmicutes	Bacteria	457570
9	Firmicutes	Bacteria	370438
d2	Firmicutes	Bacteria	138119
2P	Firmicutes	Bacteria	485916
68	Firmicutes	Bacteria	350688
12	Firmicutes	Bacteria	386415
16	Firmicutes	Bacteria	431943
0Z	Firmicutes	Bacteria	394503
3	Firmicutes	Bacteria	290402
7V	Firmicutes	Bacteria	645463
1N	Firmicutes	Bacteria	592027
ca	Firmicutes	Bacteria	272562
5E	Firmicutes	Bacteria	521460
h3	Firmicutes	Bacteria	309798
c9	Firmicutes	Bacteria	246194
m8	Firmicutes	Bacteria	264732
82	Firmicutes	Bacteria	340099
5D	Firmicutes	Bacteria	373903
ls	Firmicutes	Bacteria	314315
4L	Firmicutes	Bacteria	220668
79	Firmicutes	Bacteria	405566
ld	Firmicutes	Bacteria	390333
52	Firmicutes	Bacteria	387344
6U	Firmicutes	Bacteria	272623
4A	Firmicutes	Bacteria	543302
li	Firmicutes	Bacteria	272626
6W	Firmicutes	Bacteria	683837
oi	Firmicutes	Bacteria	221109
j6	Firmicutes	Bacteria	491915
gk	Firmicutes	Bacteria	235909
bs	Firmicutes	Bacteria	224308
lf	Firmicutes	Bacteria	279010
bh	Firmicutes	Bacteria	272558
62	Firmicutes	Bacteria	315730
ko	Firmicutes	Bacteria	527024
1P	Firmicutes	Bacteria	526976
b2	Firmicutes	Bacteria	405536
aN	Firmicutes	Bacteria	398511
x4	Firmicutes	Bacteria	66692
1V	Firmicutes	Bacteria	458233

Table A.2 (contd.)

Genome ID	*Phylum*	*Superkingdom*	*NCBI Taxonomy ID*
5S	Firmicutes	Bacteria	396513
6B	Gemmatimonadetes	Bacteria	379066
k7	Nitrospirae	Bacteria	289376
aA	Planctomycetes	Bacteria	530564
6C	Proteobacteria	Bacteria	598659
ti	Proteobacteria	Bacteria	326298
yt	Proteobacteria	Bacteria	56780
dp	Proteobacteria	Bacteria	177439
3W	Proteobacteria	Bacteria	177437
2K	Proteobacteria	Bacteria	485915
6D	Proteobacteria	Bacteria	525897
dv	Proteobacteria	Bacteria	883
0V	Proteobacteria	Bacteria	525146
p8	Proteobacteria	Bacteria	338963
jw	Proteobacteria	Bacteria	404380
gs	Proteobacteria	Bacteria	243231
gm	Proteobacteria	Bacteria	269799
60	Proteobacteria	Bacteria	448385
az	Proteobacteria	Bacteria	76114
dj	Proteobacteria	Bacteria	159087
vi	Proteobacteria	Bacteria	243365
1W	Proteobacteria	Bacteria	583345
m7	Proteobacteria	Bacteria	265072
nt	Proteobacteria	Bacteria	292415
14	Proteobacteria	Bacteria	420662
rl	Proteobacteria	Bacteria	381666
rs	Proteobacteria	Bacteria	305
b5	Proteobacteria	Bacteria	271848
3C	Proteobacteria	Bacteria	320373
b6	Proteobacteria	Bacteria	331271
30	Proteobacteria	Bacteria	269482
b4	Proteobacteria	Bacteria	266265
rx	Proteobacteria	Bacteria	338969
1	Proteobacteria	Bacteria	365044
2B	Proteobacteria	Bacteria	543728
nl	Proteobacteria	Bacteria	323848
ne	Proteobacteria	Bacteria	228410
3B	Proteobacteria	Bacteria	565050
ey	Proteobacteria	Bacteria	314225
nv	Proteobacteria	Bacteria	279238
8I	Proteobacteria	Bacteria	264203

Table A.2 (contd.)

Genome ID	*Phylum*	*Superkingdom*	*NCBI Taxonomy ID*
54	Proteobacteria	Bacteria	228405
3Q	Proteobacteria	Bacteria	272943
27	Proteobacteria	Bacteria	318586
m5	Proteobacteria	Bacteria	342108
kg	Proteobacteria	Bacteria	414684
rb	Proteobacteria	Bacteria	269796
0P	Proteobacteria	Bacteria	272568
ox	Proteobacteria	Bacteria	290633
1S	Proteobacteria	Bacteria	634457
j5	Proteobacteria	Bacteria	440085
6A	Proteobacteria	Bacteria	661410
fj	Proteobacteria	Bacteria	441620
40	Proteobacteria	Bacteria	439375
ke	Proteobacteria	Bacteria	395965
jj	Proteobacteria	Bacteria	504832
rd	Proteobacteria	Bacteria	316055
wi	Proteobacteria	Bacteria	323098
nb	Proteobacteria	Bacteria	323097
kj	Proteobacteria	Bacteria	243159
3M	Proteobacteria	Bacteria	595494
33	Proteobacteria	Bacteria	380703
vp	Proteobacteria	Bacteria	419109
65	Proteobacteria	Bacteria	410291
vb	Proteobacteria	Bacteria	216895
5O	Proteobacteria	Bacteria	417398
he	Proteobacteria	Bacteria	74109
il	Proteobacteria	Bacteria	283942
h5	Proteobacteria	Bacteria	225849
hw	Proteobacteria	Bacteria	318161
88	Proteobacteria	Bacteria	392500
lw	Proteobacteria	Bacteria	167879
px	Proteobacteria	Bacteria	342610
ha	Proteobacteria	Bacteria	326442
d0	Proteobacteria	Bacteria	203122
kb	Proteobacteria	Bacteria	314275
hc	Proteobacteria	Bacteria	349521
4H	Proteobacteria	Bacteria	523791
c8	Proteobacteria	Bacteria	290398
mt	Proteobacteria	Bacteria	243233
nr	Proteobacteria	Bacteria	323261
pq	Proteobacteria	Bacteria	384676

Table A.2 (contd.)

Genome ID	*Phylum*	*Superkingdom*	*NCBI Taxonomy ID*
ps	Proteobacteria	Bacteria	160488
4T	Proteobacteria	Bacteria	205922
pa	Proteobacteria	Bacteria	381754
ri	Proteobacteria	Bacteria	259536
pz	Proteobacteria	Bacteria	335284
tc	Proteobacteria	Bacteria	317025
83	Spirochaetes	Bacteria	456481
9C	Synergistetes	Bacteria	525903
2E	Thermotogae	Bacteria	521045
jt	Thermotogae	Bacteria	484019
63	Crenarchaeota	Archaea	453591
fd	Crenarchaeota	Archaea	490899
2	Crenarchaeota	Archaea	399549
sv	Crenarchaeota	Archaea	273063
0G	Crenarchaeota	Archaea	427318
za	Crenarchaeota	Archaea	330779
34	Crenarchaeota	Archaea	384616
11	Euryarchaeota	Archaea	349307
m9	Euryarchaeota	Archaea	259564
ma	Euryarchaeota	Archaea	188937
mz	Euryarchaeota	Archaea	192952
fu	Euryarchaeota	Archaea	269797
jx	Euryarchaeota	Archaea	521011
m6	Euryarchaeota	Archaea	323259
23	Euryarchaeota	Archaea	368407
mn	Euryarchaeota	Archaea	190192
aQ	Euryarchaeota	Archaea	589924
7W	Euryarchaeota	Archaea	572546
h6	Euryarchaeota	Archaea	523850
tk	Euryarchaeota	Archaea	69014
5V	Euryarchaeota	Archaea	593117
0X	Euryarchaeota	Archaea	604354
ph	Euryarchaeota	Archaea	70601
pb	Euryarchaeota	Archaea	272844
pu	Euryarchaeota	Archaea	186497
ta	Euryarchaeota	Archaea	273075
p3	Euryarchaeota	Archaea	263820
4X	Euryarchaeota	Archaea	485914
3R	Euryarchaeota	Archaea	519442
np	Euryarchaeota	Archaea	348780
4B	Euryarchaeota	Archaea	416348

Table A.2 (contd.)

Genome ID	*Phylum*	*Superkingdom*	*NCBI Taxonomy ID*
aG	Euryarchaeota	Archaea	309800
hm	Euryarchaeota	Archaea	272569
3P	Euryarchaeota	Archaea	573064
mj	Euryarchaeota	Archaea	243232
m2	Euryarchaeota	Archaea	444158
57	Euryarchaeota	Archaea	406327
md	Euryarchaeota	Archaea	187420
m4	Euryarchaeota	Archaea	339860
9Q	Euryarchaeota	Archaea	634498
9O	Euryarchaeota	Archaea	439481